SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Distribution limited to U.S. DOD agencies only. Operational/ Administrative information contained. Other requests for this document must be referred to the Commandant of the Marine Corps (Code RD).

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
CRC 540

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Center for Naval Analyses

6b. OFFICE SYMBOL *(If applicable)*
CNA

7a. NAME OF MONITORING ORGANIZATION
Commandant of the Marine Corps (Code RD)

6c. ADDRESS *(City, State, and ZIP Code)*
4401 Ford Avenue
Alexandria, Virginia 22302-0268

7b. ADDRESS *(City, State, and ZIP Code)*
Headquarters, Marine Corps
Washington, D.C. 20380

8a. NAME OF FUNDING / ORGANIZATION
Office of Naval Research

8b. OFFICE SYMBOL *(If applicable)*
ONR

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
N00014-83-C-0725

8c. ADDRESS *(City, State, and ZIP Code)*
800 North Quincy Street
Arlington, Virgina 22217

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
65153M	C0031		

11. TITLE *(Include Security Classification)*
Examining the Validity of Hands-On Tests as Measures of Job Performance

12. PERSONAL AUTHOR(S)
Paul W. Mayberry

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM TO

14. DATE OF REPORT *(Year, Month, Day)*
May 1986

15. PAGE COUNT
90

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
05	09	
05	10	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Aptitude tests, Enlisted personnel, Hands-on tests, Job training, JPM (Job Performance Measurement), Marine Corps, Mathematical analysis, MOS (Military Occupational Specialty), Performance (human), Performance tests, Qualifications, Skills, Statistical analysis, Validation

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

Having been designated as the benchmark for assessing on-the-job performance, hands-on tests need to be examined for the quality of their measurement. This analysis evaluates the measurement validity of hands-on tests based on the results of tests developed for three Marine Corps MOSs: Ground Radio Repair, Automotive Mechanic, and Infantry Rifleman.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Lt. Col. G.W. Russell

22b. TELEPHONE *(Include Area Code)*
(202) 694-3491

22c OFFICE SYMBOL
RDS-40

83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

CRC 540 / May 1986

EXAMINING THE VALIDITY OF HANDS-ON TESTS AS MEASURES OF JOB PERFORMANCE

Paul W. Mayberry

DISTRIBUTION STATEMENT

Distribution limited to U.S. DOD agencies only. Operational/Administrative information contained. Other requests for this document must be referred to the Commandant of the Marine Corps (Code RD).

A Division of CNA Hudson Institute

CENTER·FOR·NAVAL·ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

A Division of CNA Hudson Institute

CENTER·FOR·NAVAL·ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

30 July 1986

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Contribution 540

Encl: (1) CNA Research Contribution 540, "Examining the Validity of Hands-on Tests as Measures of Job Performance," by Paul W. Mayberry, May 1986

1. Enclosure (1) is forwarded as a matter of possible interest.

2. The armed services are engaged in a large-scale effort to use job performance to establish enlistment standards. Hands-on tests are generally considered to be the most definitive measures of performance. The purpose of this Research Contribution is to evaluate the measurement validity of several prototype hands-on tests.

3. Research Contributions are distributed for their potential value in other studies and analyses. They do not necessarily represent the opinion of the Department of the Navy.

Christopher Jehn

Christopher Jehn
Director
Marine Corps Operations
Analysis Group

Distribution List:
Reverse Page

Subj: Center for Naval Analyses Research Contribution 540

Distribution List

SNDL	
A1	ASSTSECNAV MRA
A1	DASN - MANPOWER
A2A	NAVCOMPT
A2A	OPA
A6	HQMC MPR
	Attn: Deputy Chief of Staff for Manpower (2 copies)
	Attn: Director, Personnel Procurement Division (2 copies)
	Attn: Director, Manpower Plans and Policy Division (5 copies)
	Attn: Director, Personnel Management Division (2 copies)
A6	HQMC TRNG
A6	HQMC RD&S
A6	HQMC MCH&M
B1B	ASD/FM&P
B1B	ASD/FM&P (AP)
B1B	ASD/FM&P (MM&PP)
B1B	DPAE
B2A	DTIC (2 copies)
B3	NDU
B3	AFSC
E3D1	CNR
E3D5	NAVPERSRANDCEN
	Attn: Director, Manpower and Personnel Laboratory
	Attn: Technical Library
FF38	USNA
	Attn: Nimitz Library
FF42	NAVPGSCOL
FF44	NAVWARCOL
FF48	OECEN WASHINGTON
FJA1	COMNAVMILPERSCOM
FJB1	COMNAVCRUITCOM
FT1	CNET
FT87	LDRORGEFSCOL
V12	CG MCDEC

OPNAV

OP-09BH
OP-91
OP-01
OP-11
OP-13
OP-15
OP-135L

Subj: Center for Naval Analyses Research Contribution 540

Other

Department of the Army Library
Department of the Army Headquarters (Code DAPE-MPE-CS)
Army Research Institute
 Attn: Director, Manpower and Personnel Laboratory
 Attn: Director, Personnel Utilization Technical Area
 Attn: Technical Library
Department of the Air Force (SAMI)
Department of the Air Force (AF/MPXOA)
Hq, Air Force Manpower and Personnel Center (Code MPC/YPT)
Air Force Human Resources Laboratory
 Attn: AFHRL/MOA
 Attn: AFHRL/Technical Library
Hq, Military Enlistment Processing Command (Code MEPCT-P)
Hq, U.S. Coast Guard (Code G-P-1/2/TP42)
Institute for Defense Analyses
Human Resources Research Organization
The Rand Corporation
American Institutes for Research (5 copies)
Joint Service Job Performance Measurement Working Group (12 copies)
National Academy of Sciences Advisory Committee (22 copies)

CRC 540 / May 1986

EXAMINING THE VALIDITY OF HANDS-ON TESTS AS MEASURES OF JOB PERFORMANCE

Paul W. Mayberry

Marine Corps Operations Analysis Group

A Division of CNA *Hudson Institute*

CENTER·FOR·NAVAL·ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

Having been designated as the benchmark for assessing on-the-job performance, hands-on tests need to be examined for the quality of their measurement. This analysis evaluates the measurement validity of hands-on tests based on the results of tests developed for three Marine Corps MOSs: Ground Radio Repair, Automotive Mechanic, and Infantry Rifleman.

EXECUTIVE SUMMARY

INTRODUCTION

The military services have embarked on a large-scale research project to validate enlistment standards against job performance. The service-wide project is coordinated by the Joint Services Job Performance Measurement Working Group. This group has designated hands-on tests as the benchmark measure of job performance. Within this Joint Service context, the Marine Corps conducted a study in 1981 to evaluate the feasibility of measuring job performance with hands-on tests. The results of this feasibility study were to have direct implications for the full-scale Marine Corps effort to develop and administer hands-on tests to representative military occupational specialties, beginning with the Infantry Occupational Field.

Purpose

Hands-on tests have intrinsic validity because of their high fidelity to job behavior. The validity, however, can be threatened by multiple sources of error. The purpose of this report is to:

- Present factors that affect the validity of hands-on tests as a measure of the examinees' performance on the job.
- Analyze the validity of hands-on test scores collected by the Marine Corps during the feasibility study.
- Document the implications of this analysis for the full-scale Marine Corps Job Performance Measurement Project as part of the Joint Service effort.

Components of Hands-on Test Validity

Based upon previous analyses of the feasibility study data as well as the experiences of the other services in the Joint Service arena, four principal components of hands-on measurement validity have emerged:

- Content validity. The content of hands-on tests is frequently restricted to behavior easy to test in the hands-on mode. A table of specifications, consisting of skills and knowledge on one dimension and duty areas and tasks on the other, should be used to define job requirements and select test content that represents the overall job responsibilities.

- Equivalence of test administrators. Test administrators as a rule apply idiosyncratic scoring standards. They should be trained to use the same scoring standards and then monitored continuously to make sure they remain calibrated to the standards.

- Consistency of task measures. Job tasks are organized by the Marine Corps, and each service, into occupational specialties. People in a specialty are expected to be proficient on the critical tasks. Therefore, the intercorrelation among the tasks in the test should be positive or, at the least, not negative.

- Standardized testing conditions. All examinees should be exposed to identical conditions and materials, given exactly the same directions, and allowed the same amount of time. They should be indifferent as to who administers the tests or in what order the tests are given. Any factor that introduces error into the measurement process is best controlled for by explicitly removing its effects, as in training test administrators. When these influential factors cannot be adequately controlled, it is best that they be distributed randomly across all examinees, administrators, or tasks—for example, by randomly assigning examinees to a testing order.

ANALYSIS OF HANDS-ON TEST SCORES

In the 1981 Marine Corps feasibility study, hands-on tests were developed for three Military Occupational Specialties (MOSs):

- Ground Radio Repair, with high technical requirements; 37 weeks of formal school training

- Automotive Mechanic, with moderate technical requirements; 13 weeks of formal school training

- Infantry Rifleman, with low technical requirements; 5 weeks of formal school training.

The tests were administered during the summer and fall of 1981 at Camp Pendleton, California, by Marine Corps job experts in the MOSs. The analyses of the hands-on test scores addressed the four components of hands-on measurement validity.

Content Validity

The content of these tests did not adequately represent the full range of job requirements. The problems were mainly due to:

- The lack of an explicit table of specifications to provide the blueprint for task selection. Table I compares the job requirements defined by the MOS Manual for the Ground Radio Repair specialty to the tasks that were included in the hands-on test. It is readily apparent that the single content area and single skill area did not begin to cover the diversity of the Ground Radio Repair specialty.

- Inappropriate levels of task difficulty. For two specialties (Radio Repair and Automotive Mechanic), the hands-on tests provided little useful information with respect to individual differences at the top end of the measurement scale.

Equivalence of Administrators

The analyses to evaluate the equivalence of test administrators were conducted for only two of the specialties because administrators were not identified for the Infantry Rifleman tests. In neither case were the administrators found to be equivalent raters. This degree of nonequivalence is illustrated in figure I for the Automotive Mechanic specialty. The profile shows significant discrepancies among the administrators for each of the six duty areas of the specialty. Similar findings were noted for the Ground Radio Repair specialty.

Consistency of Task Measurement

Consistency of task measurement, as defined by positive intercorrelations of the tasks, was found for each of the specialties. The correlations among the tasks of the Infantry Rifleman hands-on test are noted in table II.

Negative correlations with the "firing upon friendly targets" task are expected because of a reverse score scale.

TABLE I

COMPARISON OF JOB REQUIREMENTS FOR GROUND RADIO REPAIRERS WITH TASKS SELECTED FOR TEST

Defined in MOS Manual	Selected for test
SKILLS AND KNOWLEDGE	
Diagnose faults	Diagnose faults
Replace components	
Inspect	
Align	
Requisition parts	
Complete records	
Interconnect equipment	
CONTENT AREAS (EQUIPMENT)	
AM radios	Circuit boards
FM radios	
Terminals	
Control units	
Secure voice systems	
Multichannel radios	
Electronics items	

Source: MOS Manual

Testing Conditions

Little if any control was exercised over the testing conditions to ensure standardization in the feasibility study. Test administrators were not initially trained to rate hands-on performance, nor were their scoring techniques monitored. Likewise, randomization was not explicitly employed in any of the data collection decisions.

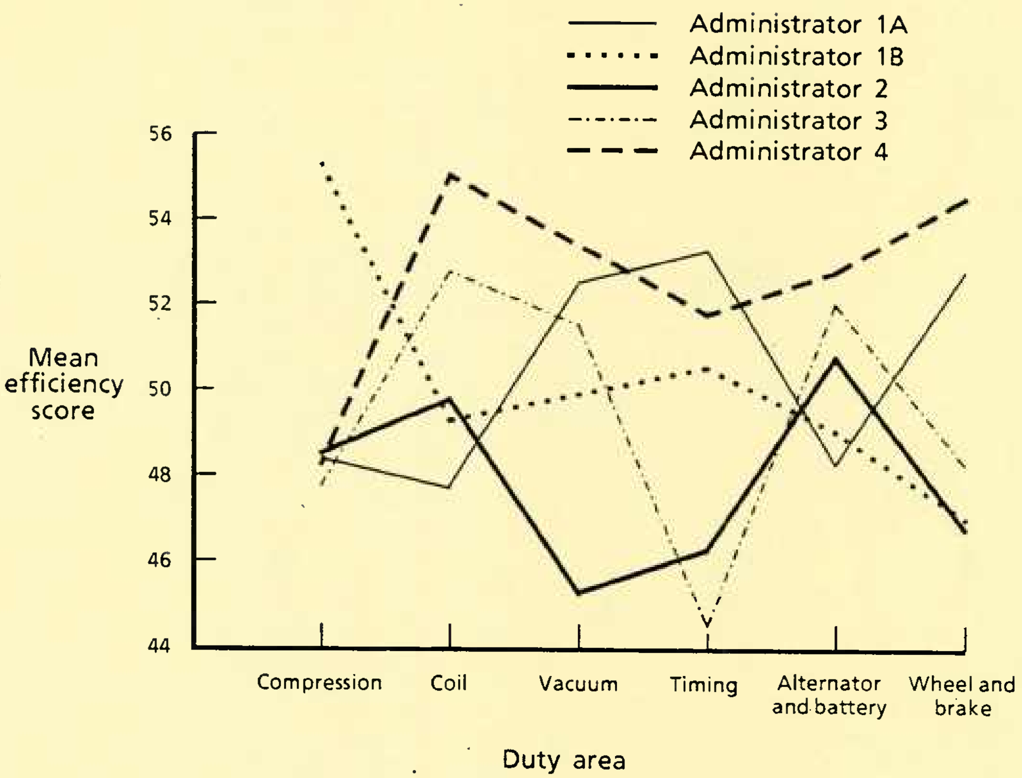

FIG. I: EQUIVALENCE OF ADMINISTRATOR PROFILES FOR AUTOMOTIVE MECHANIC SPECIALTY

IMPLICATIONS FOR THE MARINE CORPS JOB PERFORMANCE MEASUREMENT PROJECT

The analyses of the 1981 hands-on tests have resulted in improvements in the design of the Marine Corps Job Performance Measurement (JPM) Project in each of the four areas that affect measurement validity. The initial set of hands-on tests to be developed for the Marine Corps JPM Project will be for the Infantry Occupational Field.

Content Validity

A detailed table of specifications containing duty areas as one dimension, skills and knowledge underlying performance as another, and ratings of difficulty to learn as the third will guide selection of the test content. The content of the MOSs in the Infantry Occupational Field is contained in the Individual Training Standards prepared by the Marine Corps. The table of specifications permits relatively precise statements to be made about the representativeness of the test content and enhances the generalizability of test scores to performance in the MOS.

TABLE II

CORRELATIONS AMONG TASK SCORES ON THE INFANTRY RIFLEMAN HANDS-ON TEST

		Task											
Task		1	2	3	4	5	6	7	8	9	10	11	12
1	Target	1.00											
2	Friend	.06	1.00										
3	Stomach	.03	.03	1.00									
4	Jaw	.03	−.06	.21	1.00								
5	Ableed	.05	.03	.36	.13	1.00							
6	Map	.10	−.08	.13	.14	.09	1.00						
7	Compass	.03	−.14	.05	.13	.09	.44	1.00					
8	Terrain	.15	−.10	.09	.19	.10	.63	.44	1.00				
9	Symbols	.00	−.13	.04	.03	.06	.33	.22	.29	1.00			
10	Situatns	.09	−.09	.10	.03	.03	.33	.15	.28	.47	1.00		
11	Removemn	.23	−.02	.00	.14	.07	.28	.22	.30	.07	.15	1.00	
12	Armmine	.12	−.01	.14	.12	.23	.24	.22	.27	.00	.12	.31	1.00
	Mean	.55	.16	.51	.50	.59	.50	.55	.42	.59	.52	.44	.58
	Std	.23	.20	.21	.34	.18	.23	.38	.18	.22	.25	.24	.27
	N	259	259	259	259	259	259	259	259	259	259	259	259

NOTE: The correlation matrix and descriptive statistics are based on proportion-correct scores. The firing upon friendly targets scale is reversely scored. The meaning of the task labels and the respective duty area (in caps) for each task are as follows:

TARGET ENGAGEMENT		
	Target:	Target firing
	Friend:	Firing upon friendly targets
FIRST AID		
	Stomach:	Stomach wound
	Jaw:	Jaw wound
	Ableed:	Arterial bleeding
MAP AND COMPASS		
	Map:	Use of map
	Compass:	Use of compass
	Terrain:	Identify terrain features
FIRE TEAM FORMATION		
	Symbols:	Identify symbols
	Situatns:	Reaction in specific situations
ANTITANK MINES		
	Removemn:	Disarm and remove mine
	Armmine:	Arm mine

Equivalence of Administrators

The biggest impact of these analyses centers on the training and monitoring of test administrators. Profiles of scores assigned by each administrator to each examinee for each task will be used to help ensure equivalence of administrators. These profiles will show any systematic deviations among the administrators and indicate those idiosyncratic scoring standards that must be corrected. The test administrators should be thoroughly trained during the tryout of the tests, and then continue as administrators for the main sample of examinees. The scoring standards used by the administrators will be monitored daily.

Consistency of the Measures

The consistency of the measures must be evaluated during the tryout phase of the JPM project. If a measure is found to be negatively correlated with tasks, changes can be made and evaluated in a timely manner. Any changes, of course, must not degrade content validity. As a result of analyzing the hands-on tests in the feasibility study, the correlational evaluation of the consistency of the measures allows for less reliance on expert judgment and equivalence of the test administrators as the basis for inferring consistency.

Standardized Testing Conditions

The goal of valid hands-on measurement, which is that examinees are indifferent as to who tests them and when and where they are tested, can be achieved only if the testing conditions are standard for all examinees. A change resulting from these analyses is that test administrators will be randomly assigned to testing stations and rotated through all stations. Examinees will be randomly assigned to their first testing station. The daily monitoring of test administrators also helps ensure standard testing conditions.

The analyses of the hands-on tests in the feasibility study have drawn attention to multiple sources of error in hands-on measurements. Through rigorous control over the testing process, many of the problems can be avoided, resulting in valid hands-on measurements.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

SECTION 1

INTRODUCTION

The Joint Service Job Performance Measurement Working Group has established hands-on tests as the benchmark for assessing on-the-job performance. Within this Joint Service context, the Marine Corps conducted a study in 1981 to evaluate the feasibility of measuring job performance with hands-on tests as well as to examine the issue of validating aptitude test scores against job performance. The results of this feasibility study were to have direct implications for the full-scale Marine Corps effort to develop and administer hands-on tests for representative military occupational specialties (MOSs), beginning with the Infantry Occupational Field.

Being the standard against which other performance measures are to be compared, hands-on tests need to be examined for their validity of measurement to determine the extent to which:

- The tests consistently and accurately measure job performance.
- Performance on the test generalizes to performance in the job.
- A strong foundation can be established for conclusive statistical relationships.

Threats to the measurement validity of the hands-on tests abound. Such threats include:

- Nonrepresentative test content or inappropriate task difficulty
- Inconsistent scoring standards applied by test administrators
- Lack of randomization in the research design
- Inappropriate sampling or selection of examinees
- Nonstandard test administration procedures or conditions
- Inappropriate application of statistical tests.

These threats can invalidate any research findings unless steps are taken to counter or at least estimate their influence. This analysis addresses these concerns by examining the measurement validity of hands-on tests.

Based upon previous analyses of the feasibility study data as well as the experiences of the other services in the Joint Service arena, four principal components of measurement validity have emerged:

- Appropriateness of test content
- Equivalence of test administrators
- Consistency of task measurement
- Standardization of test administration procedures.

By adhering to these necessary components of measurement validity, the quality of hands-on measurement will be enhanced. The purpose of this analysis is to apply these requisite components to the data collected during the feasibility study.

The analysis is sequential, with each component building on the previous results. Hands-on tests are of little use if they do not represent what an individual is required to do on the job. The test scores would provide little information concerning an individual's ability. The same is true for the equivalence of administrators – if the administrators do not score performance equivalently, the issue of consistency of task measurement is irrelevant. Next, it is necessary that all tasks generally measure the same concept (i.e., are positively related) so that the sum of their scores reflects the job performance construct. Finally, it is essential that testing administration procedures and conditions are standardized and that the measurement process does not influence examinees' performance.

TEST CONTENT REPRESENTATIVE OF JOB REQUIREMENTS

To make accurate inferences about on-the-job performance from hands-on test performance, the test must represent the critical aspects of the job. Representativeness is facilitated by having job experts specify the job requirements and behaviors in detail. There are no commonly used procedures to empirically determine the content validity of a test other than by inspecting the items to decide if, in the aggregate, they adequately represent the job domain.

The Marine Corps has already devoted much effort to this task by developing the Individual Training Standards (ITS) for selected military occupational specialties (MOSs). The ITS provides a hierarchical statement of the standards of performance for each task within the MOS.

In addition, MOS Manuals, Programs of Instructions, and occupational surveys are available to help job experts define the requirements for a particular job. Although not as detailed as the ITS, these materials provide an overview of the necessary tasks and skills.

It is also necessary to identify specific skills and knowledge (analogous to behavioral objectives used in the development of achievement tests) that are consistent across job requirements. In this manner, generalization of performance on the hands-on test to performance in the MOS is enhanced by explicitly outlining the linkage of the test items to the domain of requirements established for the MOS at large.

By combining the job content dimensions identified by the job experts with the required skills and knowledge, a two-dimensional table of specifications is formed. Such a table provides the blueprint for test design by providing a matrix that clearly indicates the scope and emphasis of the job requirements.

The tasks to be included on the test should be selected randomly from the table of specifications at the level at which the elements are interchangeable. By carefully constructing and following the table of specifications in the building of the test, one has taken steps to ensure the test's content validity.

Individuals differ in their ability to do their jobs. Likewise, some job requirements are harder than others. It follows that if a hands-on test is a valid measure of job proficiency, then a reasonable distribution of scores would be expected, that is, not everyone would do extremely well or extremely poorly, but rather there would be a moderate spread of scores. Individual differences in proficiency would then be reflected in the variability of the hands-on test scores.

This argument about variability suggests that an additional dimension should be incorporated into the table of specifications: task difficulty. The Air

Force has done much research on this issue in identifying the "difficulty to learn" for job requirements of the Air Force specialties [2]. By ranking the tasks of each duty area (or content area) on a difficulty-to-learn scale, precaution is taken to ensure that the hands-on test is appropriate with respect to overall task difficulty within the MOS, and that test scores will be distributed according to the examinees' true job proficiency in the MOS.

Figure 1 summarizes the relationship between job content (represented by tasks within duty area), difficulty-to-learn ratings, and skills and knowledge. Given the job requirements of most MOSs, the skills and knowledge will tend to nest within duty areas, i.e., there will be few skills and knowledge elements common to all duty areas and few entries in the off diagonal cells of the figure. The tasks within each duty area are ranked by ratings of difficulty to learn. Note that the harder task tends to have more skills and knowledge elements.

Sampling of test content should be random and include all duty areas, difficulty ranges, and skills and knowledge in proportion to their frequency in the table of specifications. Such random sampling will allow for generalization of test performance across the three dimensions to the entire MOS and provide a firm basis for selecting the test content.

EQUIVALENCE OF TEST ADMINISTRATORS

It is also necessary to evaluate the equivalence of the test administrators to determine the extent to which they can *consistently* score performance. Test administrators are an integral component of hands-on assessments and, unlike paper-and-pencil tests in which scoring keys are known to be correct and consistent, administrators of hands-on tests can introduce major errors and thereby reduce the test's validity.

Administrators must be trained so that they are interchangeable and parallel, and so that examinees have no preference as to who tests them. However, hands-on tests are generally open-ended, not multiple choice. There is often a large range of possible responses, more than one correct answer, and varying degrees of response accuracy that may deserve "partial credit." The test administrator must process all of this information and produce a score that reflects the individual's ability. The administrator must make these decisions across various tasks within the test, across time, and across examinees, and they must agree with other administrators. The continuous

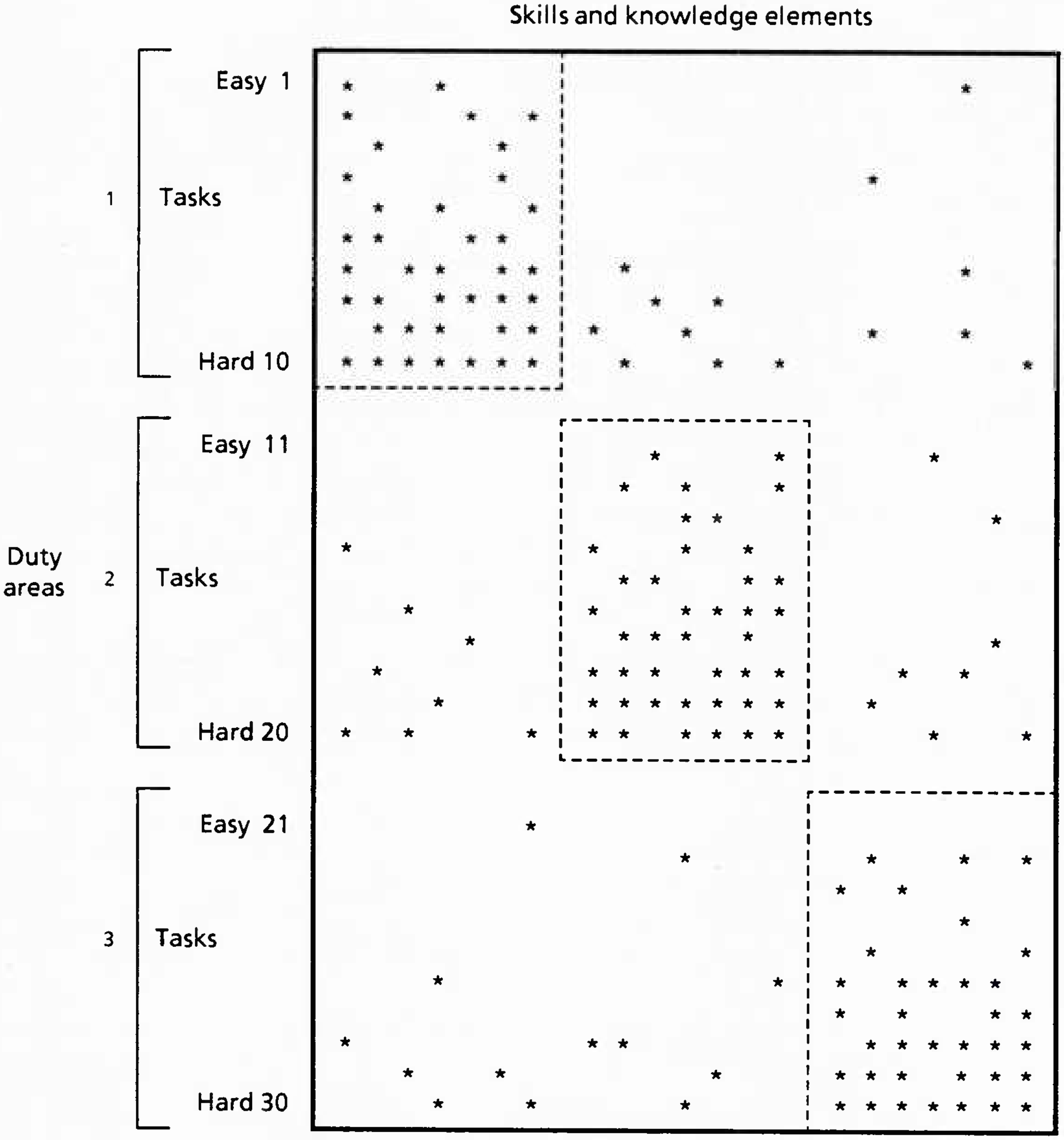

FIG. 1: TABLE OF SPECIFICATIONS INCORPORATING THREE DIMENSIONS

monitoring of administrators for equivalency is just as important as their initial training.

Scoring equivalency among administrators should result in more than just equal mean scores for the tasks. It should also result in equal distributions of scores in terms of variances and covariances across the tasks scored by each administrator. Equality of variances and covariances is a prerequisite to comparisons of mean differences. The statistical test for the equality assumption compares the covariance matrix for each administrator to a population covariance matrix (estimated by the pooled covariance matrix). (See appendix A for the specific formulas for this statistical test.)

The object of this statistical test is to determine if administrators can be pooled, i.e., if the relationship among the tasks can be expressed by one matrix for all administrators rather than a different matrix for each administrator. If the matrices are significantly different, the administrators are not applying the same scoring standards and further administrator training is required. If the matrices do not differ, the administrators are essentially parallel in scoring.

In order for the administrators to be deemed equivalent, their intercorrelations must be 0.8 or better. This threshold is restrictive enough to result in consistent measurement while still allowing for the slight influence of random error.

In summary, test administrators should be trained and monitored so that they apply the same scoring standards to result in:

- Equal means of task scores assigned by all administrators

- Equal variances and covariances of task scores assigned by all administrators

- intercorrelations among administrators of 0.8 or better.

The graphic display of figure 2 presents the conceptual framework for addressing the equivalence among the administrators. The profiles of scores assigned by each administrator are actually interaction plots of administrators and duty areas so that duty areas are on the X axis, and scores (proportion correct, for example) are on the Y axis. The lines represent the mean proportion-correct score assigned by each administrator for each duty area.

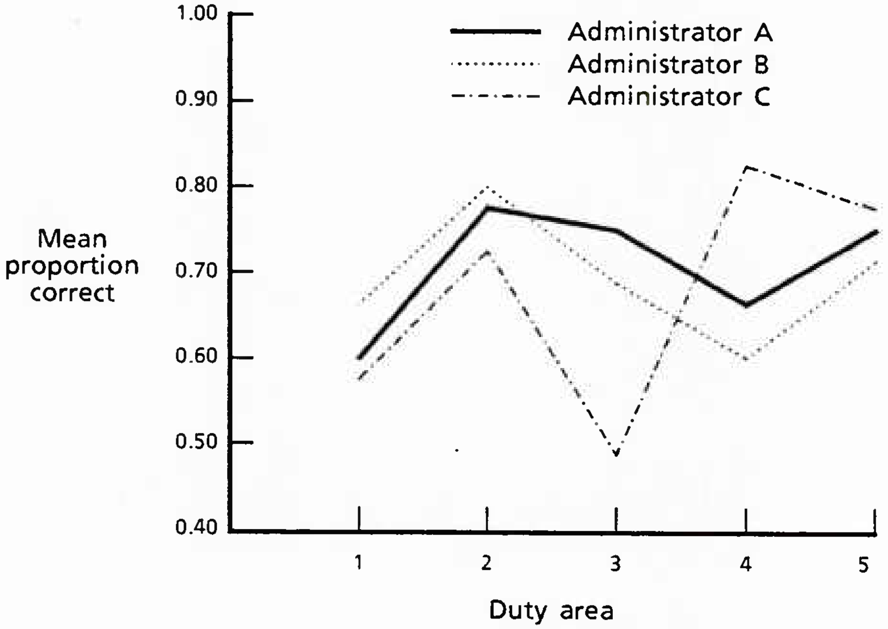

FIG. 2: EQUIVALENCE-OF-ADMINISTRATOR PROFILES

If the administrators are using the same scoring standards, the lines of the profile should be essentially parallel, be close together, and not intersect. From the figure, it is apparent that administrators A and B are basically equivalent despite the crossing of their lines between duty areas 2 and 3. For duty area 3, administrator C is much more difficult than the other administrators, while the opposite is true for duty area 4. Thus administrator C should be interviewed concerning his scoring standards for these deviant duty areas and recalibrated to the correct scale.

Such profiles can be plotted after any number of examinees are tested by two or more administrators. The administrators need not test the same examinees, although this is certainly preferable. Profiles based on three or more administrators are more informative because inconsistencies can be noted instead of significant differences between just two administrators. Such plots can be drawn after a week or even a day of testing. If problems exist, errant administrators can be corrected immediately. The profiles also provide excellent debriefing material as the administrators review their testing experiences.

These profiles are only gross approximations of the statistical tests discussed earlier. The degree of significance for the distance between the lines or their intersection is not known without completing the statistical tests. However, the profiles do provide immediate information about the potential sources of errors and allow for corrective steps to be taken. By using these plots in training administrators and then during administration of the tests, equivalence of administrators should be achieved.

CONSISTENT TASK MEASUREMENT

In developing hands-on tests, the primary objective is to measure job performance in a particular MOS. Job performance is a multifaceted concept, measured over many duty areas (e.g., map and compass, land mines, first aid, and target engagement for the Infantry Rifleman specialty). An individual may perform extremely well in one area but not be as proficient in another. Thus, the duty areas may not be measuring overall job performance, but rather many lesser dimensions of job performance.

While the tasks included in a test may be different in the sense of measuring different abilities, the covariances among the tasks should all be positive if the test is constructed carefully. Given the diverse nature of duty areas for some MOSs, the tasks selected for test content are not likely to be strictly parallel in the measurement sense. The test development process was not intended to select tasks that would result in parallel measures; rather the purpose was to proportionally sample the tasks required on the job. Therefore, the intercorrelation among the tasks is not likely to be equal or necessarily high.

However, tasks chosen for a hands-on test should be more highly related within duty area than across duty areas, with all tasks being positively related. It is the sign of the relationship, not the magnitude, that is important. Therefore, strict parallelism or symmetry of the tasks is not required; however, *consistency* of measurement among the tasks is required.[1]

A test in which the relationship between a task score and the total score is consistently negative defeats the purpose of the test by penalizing those individuals who do well most of the tasks. If the hands-on test has been found to be representative of job requirements (content validity) and can be accurately scored by administrators (equivalence of administrators), and still negative relationships exist, then the total test score misrepresents the true ability of the examinee. Therefore, with a negatively correlated task, one cannot sum across duty areas to derive a "job performance" score. Instead subscale scores need to be created to better reflect each individual's ability.

In summary, task scores are consistent measures of job proficiency if:

- Test content is representative of job requirements.
- Test administrators are equivalent.
- Positive intercorrelations exist among the tasks.

Figure 3 displays the conceptual relationship among tasks and examinees for a sample duty area. This time the focus is on the tasks rather than the administrators because the administrators have already been found to be equivalent. The profiles are for one administrator's scoring of each examinee for each task. Because of individual differences, these lines may be rather spread out, but they should not intersect to any significant degree. Note that administrator Z scored the three examinees rather consistently. From the figure, it is apparent that the ten tasks of the sample duty area are essentially

[1] The requirement of symmetry in the analysis of variance is a rather strict assumption that requires all tasks to measure the same construct in a statistical sense; i.e., variances and covariances should be equal across all tasks. The statistical test for the strict symmetry assumption addresses this question: to what extent do the magnitudes of the variances and covariances differ? The statistical test compares the pooled variance-covariance matrix (which has been determined to be equivalent across administrators) to a matrix with the average variance in the diagonal and the average covariance in the off-diagonal elements (see appendix A for details of this analysis). If the test statistic is not significant, then the tasks are thought to measure the same construct equally and the strict symmetry assumption is satisfied. However, if the test statistic is significant, there are important ramifications for the F test of main effects and interactions using the repeated measurements model. These ramifications are discussed in appendix A.

consistent because the examinee patterns are relatively horizontal and parallel. However, examinee C had problems with the fifth task. Since the other examinees did not have similar problems with this task, the inconsistency is attributed to the individual. The profiles should highlight any tasks that are not in accord with the others, and it should serve as a gauge of the administrator's consistency across examinees.

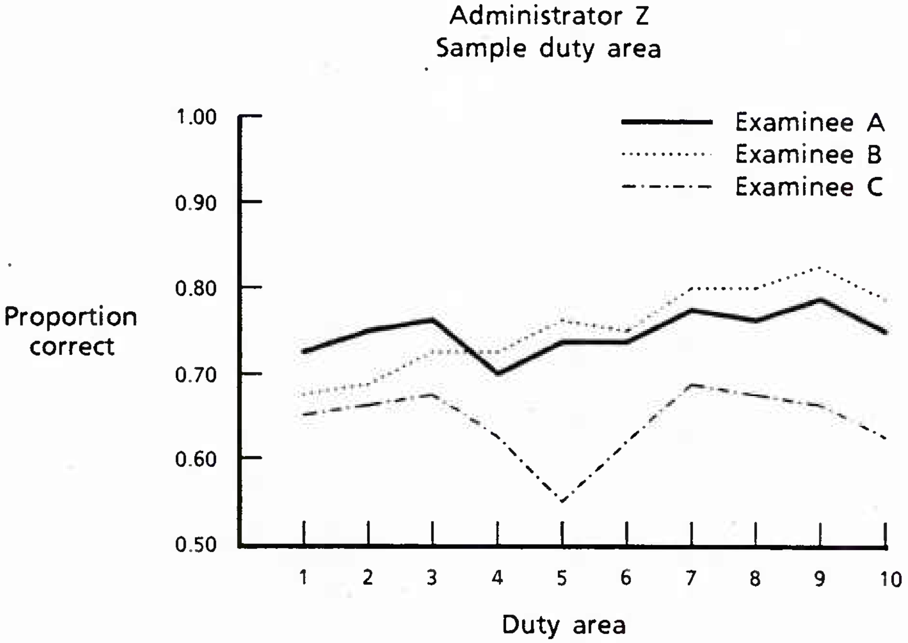

FIG. 3: SAMPLE PROFILES OF CONSISTENT TASK MEASUREMENT

The profiles can be used daily to provide the administrator with a quick evaluation of his or her performance. In the tryout stage, it is doubtful that these profiles will be very informative because the researcher will not know if the discrepancies are due to inconsistencies among the tasks, within the administrator, or between examinees. Once appropriate tasks have been selected for the full-scale administration, the profiles will be more informative. These plots may also be used to supplement the previous profiles for coaching those administrators who continuously have problems scoring the hands-on tests.

STANDARDIZATION OF TEST ADMINISTRATION PROCEDURES

The final component in establishing the measurement validity of hands-on tests involves the conditions under which the tests are administered. To ensure that the measurement process is fair and unbiased, procedures must be standardized. All examinees should be exposed to identical conditions and materials, given exactly the same directions, and allowed the same amount of time. Differences in hands-on test scores will then be a function of individual differences and not of the measurement process.

Another goal of standardized test administration is to minimize the systematic interaction among the administrators, tasks, and examinees. This is achieved by introducing randomization into the research design.

Randomization serves two important purposes. First, it controls for the potential influence of nuisance variables that are not of concern to the research. Second, it allows for generalizations of research findings beyond the specific elements of the current study to the population under consideration.

The nuisance variables include such factors as the time of day for testing, the order tests are administered, and the order of examinees. It is impossible (or infeasible) to control for all of them, but randomization can distribute the nuisance variable over all examinees, administrators, and test content. In terms of examinees, individuals should be randomly selected from among all available persons to guard against any particular group of persons being tested. Individuals should be randomly assigned to their first testing station and randomly rotated through the remaining stations to prevent testing order and time effects. Administrators should be randomly selected from the pool of available administrators. Their assignment to a testing station should be random, and likewise their rotation across stations should be random. Finally, tasks should be randomly selected from and in proportion to the frequency of equivalent items within the cells of the table of specifications.

Randomization also allows for extending and generalizing the research findings. The intent of testing people in three Marine Corps MOSs is to make inferences about the entire population of Marines within those specialties, not just those individuals who were tested. By randomly selecting and rotating administrators, inferences can also be made about all administrators, not just those used in this study. Finally, random selection of tasks for inclusion in the test ensures that generalizations can be made to all tasks of the MOS.

SECTION 2

RESULTS

The four components of measurement validity are examined in light of data collected during the feasibility study of job performance measurement conducted by the Marine Corps in 1981. The feasibility study involved three selected MOSs: Ground Radio Repair, Automotive Mechanic, and Infantry Rifleman.

SELECTING TEST CONTENT TO BE REPRESENTATIVE OF JOB REQUIREMENTS

Determining the extent to which a hands-on test is representative of job requirements is the first step in assessing its measurement validity. This subsection focuses on the process used in designing the hands-on tests as well as the implicit purpose of the tests created by each group of experts. The process of developing a table of specifications is also examined for each specialty, and the extent to which the tests adhere to officially established job requirements is discussed. Finally, the issue of appropriate test difficulty is considered.

Hands-on Test Development

The hands-on tests were developed by testing experts from the Navy Personnel Research and Development Center (NPRDC) in conjunction with Marine Corps job experts from Camp Pendleton, California. The test for each specialty was developed by a different team of experts. The test development procedures and content are described in an NPRDC report [3], which is summarized in the following discussion.

The approach involved:

- Identifying the critical tasks in each specialty.
- Ranking the task areas according to their ability to predict overall performance in the specialty.

- Constructing hands-on tests for the top-ranked tasks that are suitable for such testing.

- Conducting field tryouts of the tests.

Ground Radio Repair

Eight Marine Corps job experts reviewed the task areas for the Ground Radio Repair skill. The consensus was that troubleshooting would be the most predictive task area for the specialty. Because most circuits are similar, the specific equipment used was not considered important. To minimize the influence of experience, a unique piece of equipment that had not yet been issued to the field was chosen. The test consisted of troubleshooting ten circuit boards. A total of 210 minutes, with up to 30 minutes for each board, was allowed. For each board, the examinees were instructed to identify the symptom (worth 2 points), the faulty circuit (up to 4 points), and the faulty component (up to 8 points). Examinees were allowed to consult technical manuals and troubleshooting charts throughout the test. They were encouraged to guess when they had narrowed the choice of circuits and components. No feedback was given concerning the accuracy of their responses because the scoring rules involved a penalty for progressing from a correct to incorrect response.

The test was tried out on five Marine Corps ground radio repairers. Afterwards, the team of experts was satisfied with the test content, the scorability of the tasks, and the levels of performance by the examinees.

Automotive Mechanic

Initial efforts to identify task areas for the Automotive Mechanic specialty covered a broad scope of job requirements, including vehicle recovery, electrical systems, and intermediate-level maintenance, as well as organizational-level requirements. Subsequently, the decision was made to focus on organizational-level mechanics. Another set of job experts assigned to test development decided that major engine tune-up would be the best predictor of proficiency for automotive mechanics at the organizational level. The tune-up tasks were supplemented with tasks from wheel and brake maintenance as well as tasks for alternator and battery output.

The hands-on test consisted of 81 steps. For each examinee, the test administrator recorded either a pass or fail for each step. The time required to complete the tasks was also recorded, which permitted efficiency scores to be computed. Each examinee filled out a record of the maintenance performed during the test, called the "Equipment Repair Order." The test was tried on six examinees. The expert team felt confident that the test could be reliably scored.

During the feasibility study a misprint was noted in the scoring document for the plugs task of the engine tune-up duty area, so it was not known whether the examinees completed the task successfully. Accordingly, the plugs task is not included in this analysis.

Infantry Rifleman

For the Infantry Rifleman specialty, the hands-on test was intended to parallel combat conditions as much as possible, but to avoid conditions that could cause injury. The test included five duty areas and required about 4 hours to complete.

The test was administered to a sample of 81 examinees. Based upon this tryout, a final version was developed for full-scale administration.

In the feasibility study, testing stations representing the different duty areas were set up, and examinees rotated through them. Test administrators tended to remain at particular stations, although there were some changes. Test administrators scored the steps of each task on a pass-fail basis, including negative points for serious errors (e.g., firing on friendly targets, inability to locate north by reading a compass). A high level of agreement in scoring the tests was noted even though some administrators did not strictly conform to standard procedures.

A new set of administrators administered the tests during the full-scale testing. Although they were experts in the test content, they received little or no training in administering the test.

Purpose of Testing

While a strategy for test development had been outlined for all specialties, each panel of experts employed a different approach and therefore addressed a slightly different testing purpose. Such procedures do not necessarily lead to the same level of content validity or generalizability for the specialty.

For the Ground Radio Repair specialty, a novel approach was taken. Examinees were to apply their troubleshooting skills learned from other devices to a new piece of equipment. The test results should therefore generalize to the troubleshooting skill, but not necessarily reflect how well the examinees perform their current duties. The infantry test was composed of tasks that assessed examinees' ability to complete job requirements as well as measured their job knowledge. The Automotive Mechanic test was probably the test most true to the original purpose of hands-on testing. Tasks were chosen from those that are actually performed on the job.

A lesson learned is that during test development, the purpose of the test must be stated explicitly and the implications of any deviations must be thoroughly examined.

Table of Specifications

Although job experts attempted to organize the job requirements in each specialty prior to selecting test content, no tables of specifications were explicitly developed. Individual Training Standards (ITS) were not available at the time, so the experts had to rely on task lists, their own experiences, and brief job summaries in Marine Corps MOS Manuals. As a result, the experts differed widely about the job requirements. The development of ITS since then has formalized the process of defining specific job requirements.

Traditionally in test development two dimensions are used when preparing a table of specifications: skills and knowledge that underlie performance, and content areas. Table 1 outlines a table of specifications for the Ground Radio Repair specialty. The table compares the job requirements designated in the MOS Manual to those selected for the hands-on test. It is apparent that the test does not adequately sample the variety of tasks and skills required to be a competent radio repairer. Therefore, this hands-on test is questionable in terms of its content validity. Accordingly, performance on the test is limited in its generalizability.

TABLE 1

COMPARISON OF JOB REQUIREMENTS FOR GROUND RADIO REPAIRERS WITH TASKS SELECTED FOR TEST

Defined in MOS Manual	Selected for test
SKILLS AND KNOWLEDGE	
Diagnose faults	Diagnose faults
Replace components	
Inspect	
Align	
Requisition parts	
Complete records	
Interconnect equipment	
CONTENT AREAS (EQUIPMENT)	
AM radios	Circuit boards
FM radios	
Terminals	
Control units	
Secure voice systems	
Multichannel radios	
Electronics items	

Similar results are found for the Automotive Mechanic specialty, as shown in table 2. In the skills and knowledge dimension, the job requirements selected for the test are somewhat more representative of those defined in the MOS Manual. But again, the content area is represented by only a single element, the M-151 quarter-ton jeep. Automotive mechanics service a wide variety of vehicles, ranging from multifuel trucks to all-terrain vehicles. Competence in maintaining the M-151 does not necessarily imply competence in repairing the others.

Finally, the content areas for the Infantry Rifleman specialty were identified from a preliminary version of the Individual Training Standards (ITS), although the expert panel did not have access to the document. Note in table 3 that the ITS does not identify skill and knowledge requirements for this specialty, but it does identify 13 content areas. The hands-on test included only five of these content areas, but more importantly, the selections were not in proportion to the number of tasks in each content area. The hands-on test overemphasized first aid and land navigation while undersampling tactical measures and completely ignoring communications. As in the other specialties, the hands-on test did not adequately represent the job requirements.

TABLE 2

COMPARISON OF JOB REQUIREMENTS FOR AUTOMOTIVE MECHANICS WITH TASKS SELECTED FOR TEST

Defined in MOS Manual	Selected for test
SKILLS AND KNOWLEDGE	
Repair vehicles to second echelon	Major engine tuneup
Perform prefording and postfording maintenance	Repair wheel and brake
Complete forms	Adjust alternator
Use common repairshop equipment	Monitor battery output
Maintain testing and diagnostic equipment	Complete forms
CONTENT AREAS (EQUIPMENT)	
Vehicles up to 5-ton capacity	M-151
M-35	
M-54	
M-151	
M-561	
M-813	
M-880	

In summary, the hands-on tests for the three MOSs lack adequate content validity. (However, the panels of test development experts did not have the benefit of the ITS to use in selecting of test content.) The key to selecting content is to develop the table of specifications properly and then *randomly* sample the test elements from the table on the level at which all elements have been judged to be interchangeable.

TABLE 3

COMPARISON OF JOB REQUIREMENTS DEFINED FOR INFANTRY RIFLEMAN WITH TASKS SELECTED FOR TEST

SKILLS AND KNOWLEDGE

Not identified in Individual Training Standards.

CONTENT AREAS

	Number of tasks	
Content area	Defined in MOS	Selected for test
M16A2 service rifle	2	1
Mines	5	2
First aid	2	3
Land navigation	3	3
Tactical measures	11	2
Security/intelligence	4	–
Night-vision devices	2	–
Grenade launcher	5	–
Squad automatic weapon	3	–
Light antitank weapon	3	–
Hand grenades	3	–
Communications	10	–
Nuclear, biological, chemical warfare	5	–

Source: Individual Training Standards

Level of Task Difficulty

The test development panels did not make explicit judgments of the difficulty of the tasks they selected for inclusion in the test. Thus, while the selected tasks may represent "things that everyone should be able to do," they may be trivial requirements and not reflect the entire range of task difficulty for that specialty.

The score distributions are now examined for each specialty to note any anomalies.

Automotive Mechanic

The tasks means and standard deviations for automotive mechanics are presented in table 4. The score scale is the proportion correct. Note that administrator 1 is divided into 1A and 1B. It was learned that this administrator was providing too much feedback to examinees, so changes in his scoring strategy were imposed. The scores under 1B are those given after the changes.

TABLE 4

MEANS AND STANDARD DEVIATIONS FOR PROPORTION-CORRECT SCORES ON AUTOMOTIVE MECHANIC TEST

	Administrator				
Task	1A	1B	2	3	4
Compression	.96	.98	.88	.98	.94
	(.11)	(.07)	(.10)	(.05)	(.08)
Coil	.97	.94	.90	.97	.95
	(.14)	(.11)	(.11)	(.05)	(.15)
Vacuum	.96	.98	.91	.96	.96
	(.15)	(.07)	(.18)	(.13)	(.19)
Timing	.95	.93	.90	.99	1.00
	(.19)	(.17)	(.17)	(.04)	(.00)
Alternator	.96	.92	.88	.99	.95
	(.11)	(.18)	(.17)	(.03)	(.14)
Battery	.89	.75	.68	.85	.80
	(.15)	(.19)	(.19)	(.20)	(.09)
Wheel and brake	.97	.89	.80	.91	.98
	(.14)	(.14)	(.21)	(.27)	(.04)
Equipment repair order	.84	.83	.83	.79	.81
	(.13)	(.08)	(.08)	(.08)	(.16)

Note: Standard deviations are in parenthesis under the mean proportion-correct score for each task.

The striking conclusion drawn from table 4 is that essentially everyone received nearly perfect scores for all tasks. The median proportion correct for all tasks was 0.93. In general, administrator 2 tended to be the hardest scorer, while administrator 3 was the easiest. The vacuum and timing tasks were performed almost perfectly by everyone, while completing the equipment repair order posed the most challenge. Therefore the test did little to differentiate individual differences because of this ceiling effect – the concentration of test scores near the top of the scale.

The ceiling effect can be illustrated by profiles showing the distribution of test scores. Called box-and-whisker plots, these profiles summarize the information presented in table 4. An example of a box-and-whisker plot is presented in figure 4. The median test score for administrator Z is represented by an asterisk (*), with the box enclosing the range of scores forming the 25th to 75th percentile. Whiskers extend from the box to include scores within 1.5 interquartile ranges (one interquartile range is the distance between the 25th and 75th percentile) of the sample median. Individual scores above 1.5 but no more than 3 interquartile ranges are designated as outliers (O), and those extending beyond 3 interquartile ranges are classified as extreme cases (E). By examining these plots, one can picture the distribution of test scores and note the possible influence of deviant cases on the values of the sample standard deviation and mean.

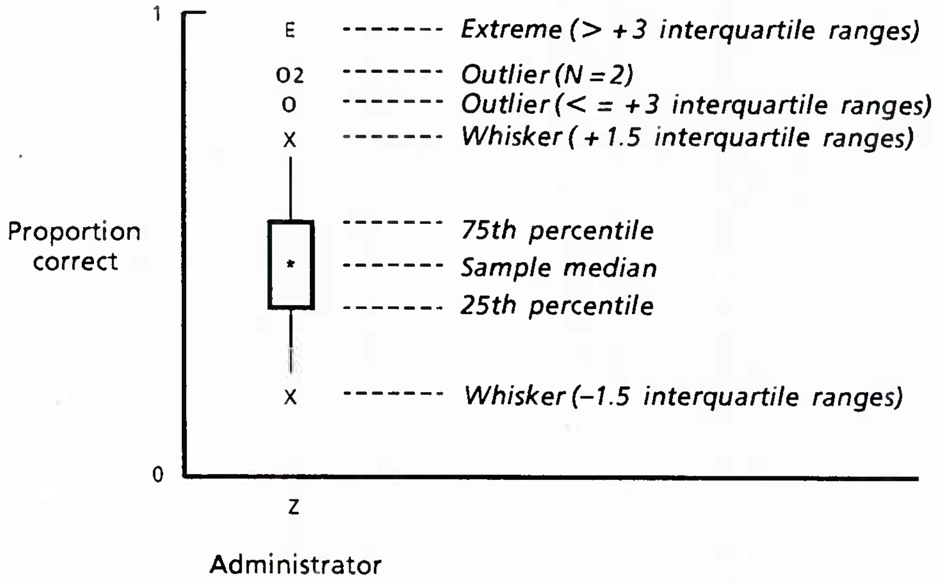
1
E ------- Extreme (> +3 interquartile ranges)
O2 ------- Outlier (N = 2)
O ------- Outlier (< = +3 interquartile ranges)
X ------- Whisker (+ 1.5 interquartile ranges)
------- 75th percentile
* ------- Sample median
------- 25th percentile
X ------- Whisker (–1.5 interquartile ranges)
Proportion correct
0
Z
Administrator

FIG. 4: EXAMPLE OF BOX-AND-WHISKER PLOT

The box-and-whisker plots for each administrator of the mechanics hands-on test are presented in figure 5. Note that for most of the tasks, the sample medians are at the top of the scale, with little or no variance among the scores. It is apparent that the outlying and extreme scores contribute most of the weight to the observed sample variances. The equipment repair order task has the most symmetric distribution of scores across all administrators, while administrator 2 is the only one to consistently score performance on all tasks, except the battery task, so as to scatter individuals over the score scale. The plots in figure 5 show that due to ceiling effects, score differences are not meaningful. Therefore in terms of proportion-correct scores, the hands-on test did not represent the full range of task difficulty in the MOS and was not sensitive enough to detect the full range of individual differences in proficiency.

In an effort to counter the lack of variance within the proportion-correct scores and to provide more information about individuals' abilities, efficiency scores were used as the unit of measurement. Efficiency scores were calculated as unit of performance per unit of time. It was necessary to combine the alternator and battery tasks because only one time measurement was made for both tasks. Also, equipment repair order was deleted because it was part of the other tasks and had no time component.

The means and standard deviations for the efficiency scores of each task are presented in table 5. While the efficiency scores do not have meaning in themselves, the transformation did improve the overall variance of the scores. Thus, by giving those individuals who score well in a short period of time more credit than those who do equally well but require more time, the score distributions are no longer plagued by the ceiling effect.

The box-and-whisker plots for the efficiency scores are presented in figure 6. The efficiency scores have more symmetric score distributions. Note that while outlying and extreme values still exist after the transformation, they are now less prevalent. The observed variances for the efficiency scores are much larger than those for the proportion-correct scores. The ceiling effect was removed, and the variance among the scores was greatly increased.

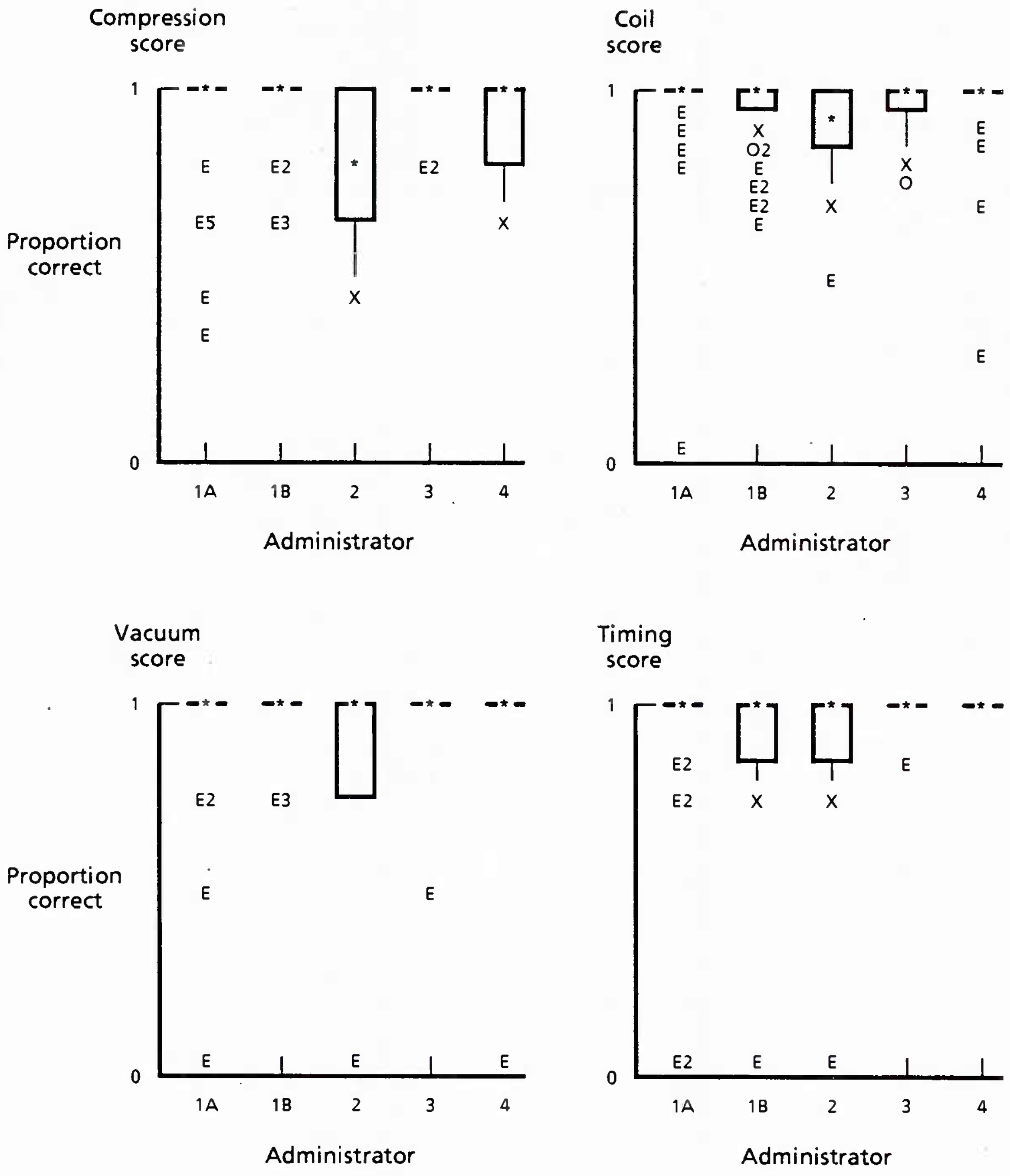

FIG. 5: BOX-AND-WHISKER PLOTS FOR PROPORTION-CORRECT SCORES IN AUTOMOTIVE MECHANIC TEST

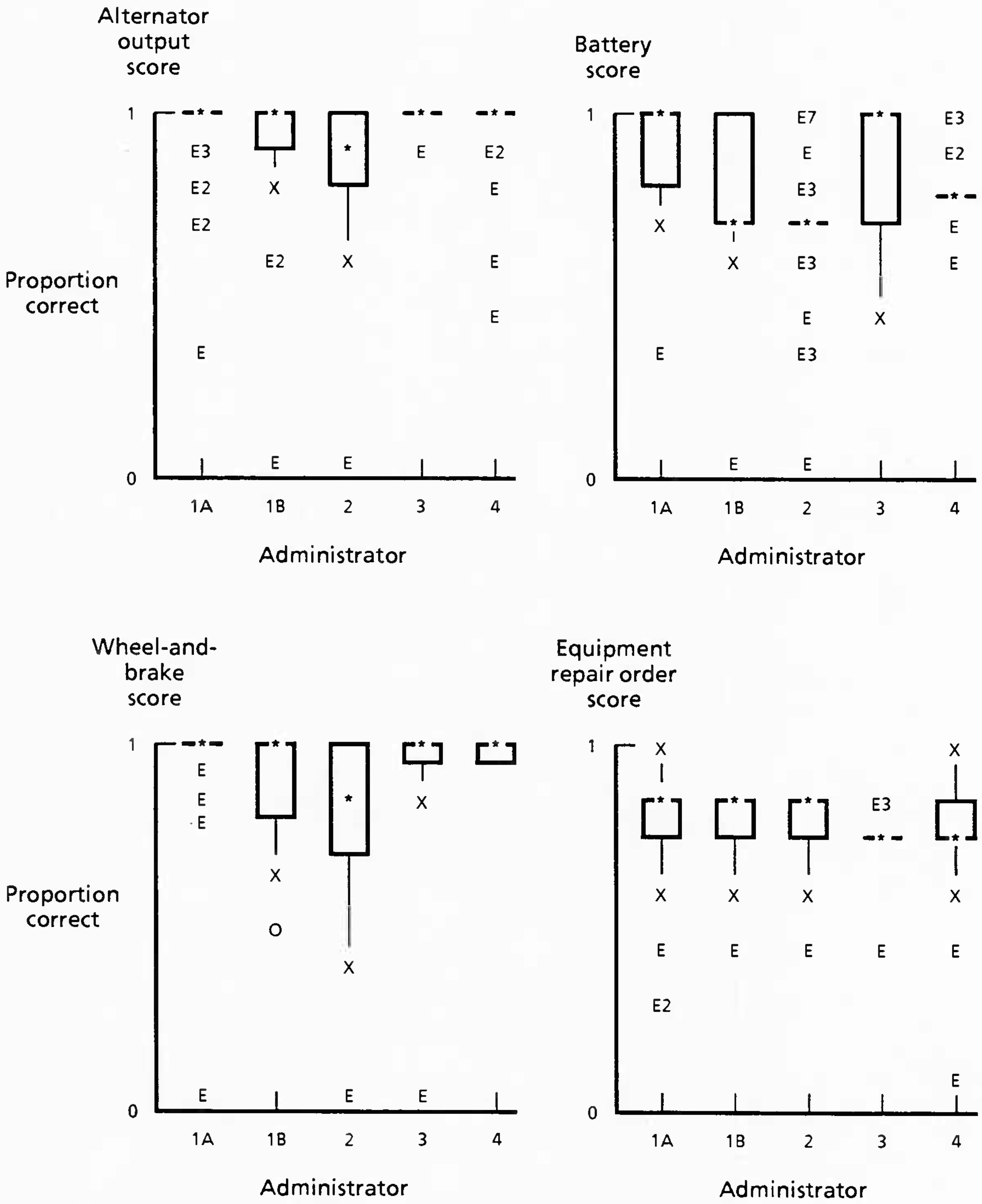

FIG. 5: (Continued)

TABLE 5

MEANS AND STANDARD DEVIATIONS FOR EFFICIENCY SCORES ON AUTOMOTIVE MECHANIC TEST

	Administrator				
Task	1A	1B	2	3	4
Compression	3.4 (1.3)	4.2 (1.4)	3.3 (1.0)	3.3 (1.2)	3.3 (1.4)
Coil	4.0 (1.5)	4.3 (1.4)	4.3 (1.6)	5.1 (1.5)	5.3 (2.2)
Vacuum	12.0 (6.1)	10.4 (3.9)	7.9 (3.6)	10.4 (5.7)	12.4 (7.1)
Timing	13.1 (6.2)	11.1 (6.7)	9.1 (5.1)	7.7 (2.9)	12.4 (5.2)
Alternator and battery	7.2 (2.2)	7.4 (2.1)	7.7 (2.2)	8.3 (5.6)	8.4 (2.7)
Wheel and brake	3.7 (2.4)	2.7 (.9)	2.7 (1.2)	2.9 (1.2)	3.9 (1.2)

Note: Standard deviations are in parenthesis under the mean efficiency score for each task. The alternator and battery tasks were combined because only one time measurement was taken for both. Equipment repair order was deleted because it had no time component.

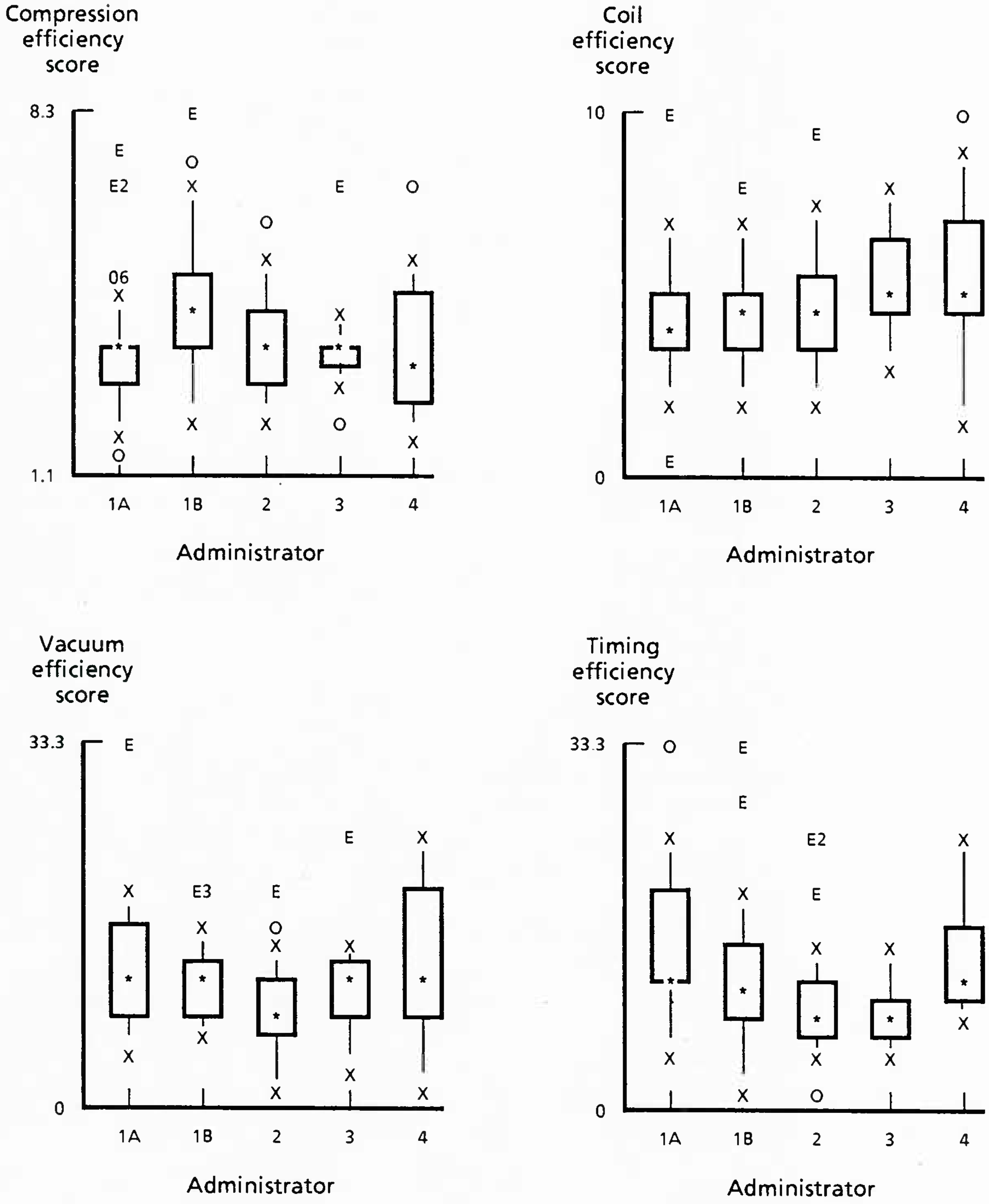

FIG. 6: BOX-AND-WHISKER PLOTS FOR EFFICIENCY SCORES IN AUTOMOTIVE MECHANIC TEST

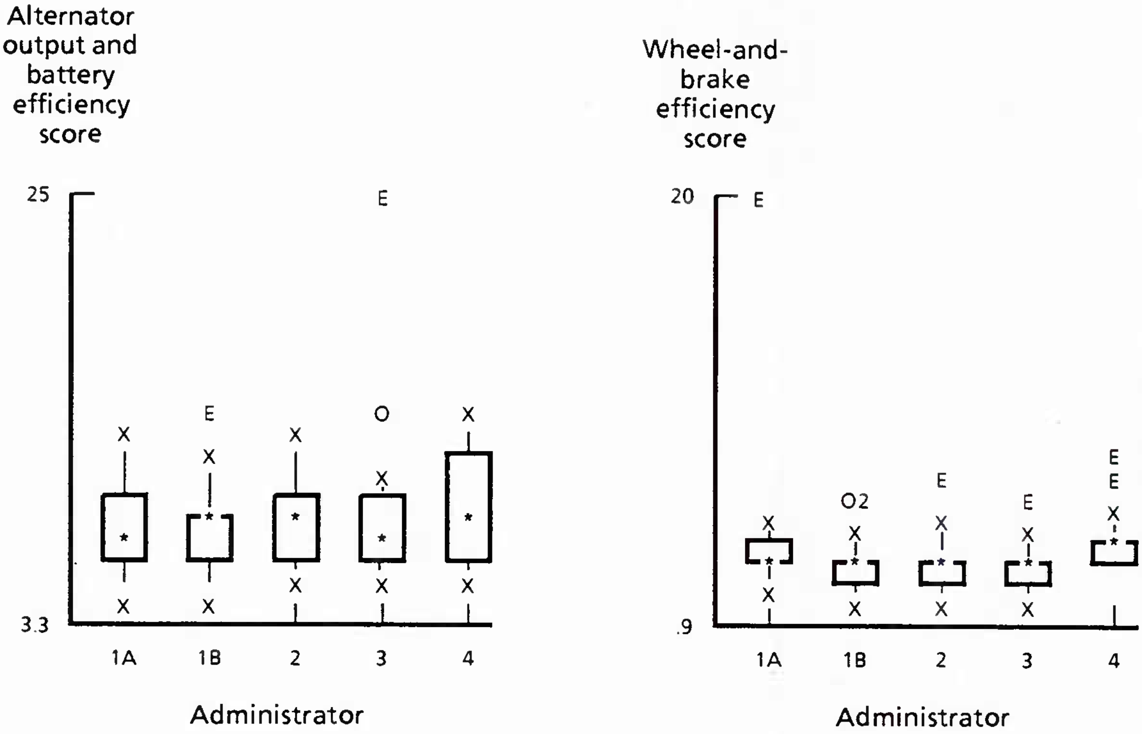

FIG. 6: (Continued)

Ground Radio Repair

The experts working on the Ground Radio Repair test likewise did not sample tasks from the entire range of task difficulty. The resulting score distributions were acceptable although still on the high side. The means and standard deviations for the ten circuit boards are given in table 6.

The important point to note is that administrators one through five differed as to who was the hardest or easiest rater depending upon the particular circuit board; however, the unknown administrators were consistently the hardest. This classification of unknown was given to those administrators who tested the first 46 examinees. These administrators were not originally required to sign the answer sheets, but the policy was changed so that administrator differences could be examined.

It is apparent that the administrators became more lenient in their scoring after this change. Such noticeable changes in scores have serious implications for the training and monitoring of administrators. Administrators should be trained to be impartial raters of ability and be aware of the feedback that they will receive.

TABLE 6

MEANS AND STANDARD DEVIATIONS FOR PROPORTION-CORRECT SCORES ON GROUND RADIO REPAIR TEST

	Administrator					
Board	1	2	3	4	5	Unknown
1	.85 (.23)	.95 (.20)	.98 (.05)	.94 (.14)	.95 (.14)	.84 (.25)
2	.86 (.31)	.89 (.29)	.98 (.07)	.89 (.19)	.84 (.29)	.85 (.25)
3	.77 (.29)	.78 (.26)	.80 (.21)	.84 (.24)	.76 (.31)	.66 (.35)
4	.84 (.35)	.93 (.14)	.91 (.18)	.92 (.24)	.69 (.40)	.74 (.33)
5	.88 (.23)	1.00 (.00)	.92 (.17)	.87 (.22)	.89 (.22)	.72 (.30)
6	.78 (.25)	.87 (.22)	.74 (.30)	.70 (.22)	.79 (.28)	.75 (.29)
7	.98 (.05)	.98 (.08)	.76 (.33)	.80 (.30)	.76 (.33)	.71 (.34)
8	.87 (.30)	.91 (.24)	.74 (.32)	.71 (.35)	.59 (.38)	.57 (.40)
9	.88 (.32)	.98 (.08)	.90 (.29)	.86 (.24)	.81 (.35)	.78 (.35)
10	.86 (.28)	.98 (.08)	.79 (.31)	.82 (.29)	.57 (.42)	.56 (.42)

Note: Standard deviations are in parenthesis under the mean proportion-correct score for each board. Adminstrators who did not originally sign the answer sheets are designated as unknown.

Infantry Rifleman

The score distributions for the Infantry Rifleman test showed much variance about the sample median, with no ceiling or floor effects. The box-and-whisker plots for proportion-correct scores are presented in figure 7. Note how the scores for each duty area cover a large range, with the boxes showing good distance between the 25th and 75th percentiles. The overall difficulty level for each duty area is approximately 0.5. Thus, it appears that this test did cover a reasonable range of task difficulty.

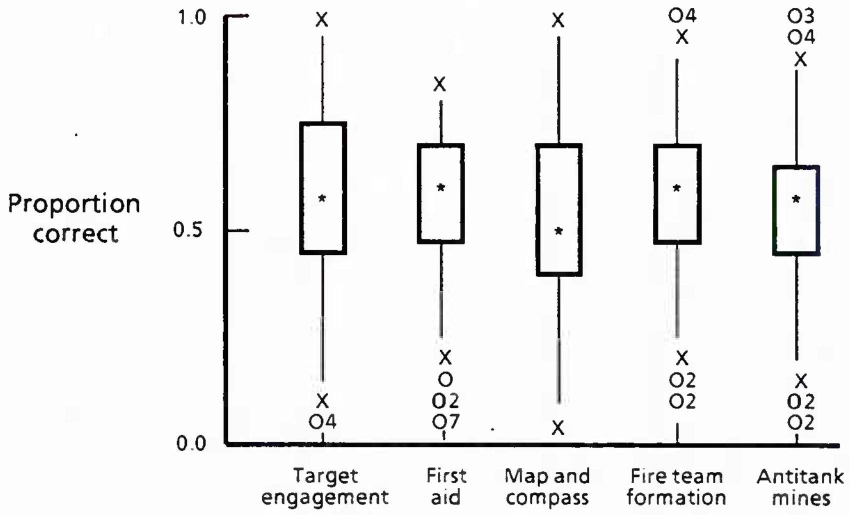

FIG. 7: BOX-AND-WHISKER PLOTS OF PROPORTION-CORRECT SCORES FOR INFANTRY RIFLEMAN TEST

EQUIVALENCE OF TEST ADMINISTRATORS

The statistical test for determining the degree of consistency among administrators was discussed earlier and is also presented in appendix A. Because administrators' names were recorded only for the Ground Radio Repair and Automotive Mechanic specialties, no results are presented for the Infantry Rifleman specialty

The results obtained in this study represent post hoc analyses, and thus administrators did not have the feedback that is the ultimate purpose of such quality control measures. Therefore, the results are only illustrative of applying statistical quality control steps, with the data collection process not benefiting from their findings.

Data editing and standardization procedures were applied to determine if a data set could be produced in which the administrators consistently scored the tasks. The steps were as follows:

Step	Rationale
1. Delete irregular discharges.	Delete individuals not typical of overall military sample.
2. Standardize duty areas, delete outliers (± 3 SD).	Equate differences in the score scale for the duty areas.
3. Log transform score scale,standardize and delete outliers (± 3 SD) (required only for efficiency scores).	Reduce proportionality of means and standard deviations; outliers overly contribute to variance.
4. Standardize administrators and duty areas, delete outliers (± 3 SD).	Control for inconsistent scoring by administrator across duty areas.
5. Delete deviant administrators.	Reduce error introduced by deviant administrators.
6. Delete deviant duty areas.	Reduce dimensionality of hands-on measure.

The steps progress from the reasonable to the extreme. If preliminary quality control measures are effective, such manipulation of the data should not be necessary. The deletion of duty areas or administrators is not advocated because the original intent of the hands-on measure would be severely changed.

Automotive Mechanic

Table 7 presents the covariance and correlation matrices for the duty areas of the Automotive Mechanic test. The efficiency scales were standardized because they were based on different time intervals and numbers of tasks for each duty area.

TABLE 7

COVARIANCE AND CORRELATION MATRICES FOR DUTY AREAS OF AUTOMOTIVE MECHANIC TEST

Total Sample

	Duty Area					
	1	2	3	4	5	6
			Panel A: Covariance Matrix			
1. Compression	100.00					
2. Coil	13.96	100.00				
3. Vacuum	16.81	7.51	100.00			
4. Timing	19.37	7.70	27.00	100.00		
5. Alt. & Battery	7.42	20.63	18.86	2.11	100.00	
6. Wheel & Brake	3.73	8.20	12.61	8.54	10.81	100.00
Mean	50.00	50.00	50.00	50.00	50.00	50.00
Std	10.00	10.00	10.00	10.00	10.00	10.00
N	176	176	176	176	176	176
			Panel B: Correlation Matrix			
1. Compression	1.00					
2. Coil	.14	1.00				
3. Vacuum	.17	.07	1.00			
4. Timing	.19	.08	.27	1.00		
5. Alt. & Battery	.07	.21	.19	.02	1.00	
6. Wheel & Brake	.04	.08	.12	.08	.11	1.00

Note: The covariance and correlation matrices are based on standardized efficiency scores for each of the duty areas.

TABLE 7 (Continued)

Administrator 1A

	Duty Area					
	1	2	3	4	5	6
Panel A: Covariance Matrix						
1. Compression	103.63					
2. Coil	23.56	80.37				
3. Vacuum	30.00	17.34	126.29			
4. Timing	35.45	15.88	49.19	110.54		
5. Alt. & Battery	20.87	11.74	29.47	23.46	70.05	
6. Wheel & Brake	12.49	−7.15	5.86	−1.02	13.47	197.70
Mean	48.57	47.66	52.65	52.97	48.26	52.76
Std	10.18	8.96	11.24	10.51	8.37	14.06
N	55	55	55	55	55	55
Panel B: Correlation Matrix						
1. Compression	1.00					
2. Coil	.26	1.00				
3. Vacuum	.27	.17	1.00			
4. Timing	.33	.17	.42	1.00		
5. Alt. & Battery	.25	.16	.31	.26	1.00	
6. Wheel & Brake	.09	−.06	.04	−.01	.12	1.00

Note: The covariance and correlation matrices are based on standardized efficiency scores for each of the duty areas.

TABLE 7 (Continued)

Administrator 1B

	Duty Area					
	1	2	3	4	5	6
			Panel A: Covariance Matrix			
1. Compression	112.46					
2. Coil	−14.60	68.79				
3. Vacuum	13.00	−5.06	49.64			
4. Timing	−9.57	10.92	−25.50	137.24		
5. Alt. & Battery	30.75	15.54	4.21	−3.83	62.96	
6. Wheel & Brake	13.22	−1.38	−1.06	9.85	−2.10	27.18
Mean	55.27	49.22	49.91	50.65	49.11	47.02
Std	10.60	8.29	7.05	11.71	7.93	5.21
N	40	40	40	40	40	40
			Panel B: Correlation Matrix			
1. Compression	1.00					
2. Coil	−.17	1.00				
3. Vacuum	.17	−.09	1.00			
4. Timing	−.08	.11	−.31	1.00		
5. Alt. & Battery	.37	.24	.08	−.04	1.00	
6. Wheel & Brake	.24	−.03	−.03	.17	−.05	1.00

Note: The covariance and correlation matrices are based on standardized efficiency scores for each of the duty areas.

TABLE 7 (Continued)

Administrator 2

	Duty Area					
	1	2	3	4	5	6
			Panel A: Covariance Matrix			
1. Compression	57.62					
2. Coil	15.24	93.63				
3. Vacuum	−4.58	7.24	43.29			
4. Timing	20.30	3.09	6.28	64.23		
5. Alt. & Battery	−19.50	24.77	14.39	−3.24	72.40	
6. Wheel & Brake	−3.80	−1.60	−1.61	.02	10.91	49.98
Mean	48.76	49.65	45.14	46.43	50.63	46.82
Std	7.59	9.68	6.58	8.01	8.51	7.07
N	41	41	41	41	41	41
			Panel B: Correlation Matrix			
1. Compression	1.00					
2. Coil	.21	1.00				
3. Vacuum	−.09	.11	1.00			
4. Timing	.33	.04	.12	1.00		
5. Alt. & Battery	−.30	.29	.26	−.05	1.00	
6. Wheel & Brake	−.07	−.02	−.03	.00	.18	1.00

Note: The covariance and correlation matrices are based on standardized efficiency scores for each of the duty areas.

TABLE 7 (Continued)

Administrator 3

	Duty Area					
	1	2	3	4	5	6
Panel A: Covariance Matrix						
1. Compression	90.00					
2. Coil	49.65	92.98				
3. Vacuum	−33.72	−10.91	104.92			
4. Timing	−16.47	−2.04	23.51	22.87		
5. Alt. & Battery	33.51	35.76	23.66	−7.24	429.93	
6. Wheel & Brake	7.60	36.51	8.65	−4.97	37.79	48.43
Mean	47.80	52.87	49.40	44.56	52.06	47.99
Std	9.49	9.64	10.24	4.78	20.73	6.96
N	13	13	13	13	13	13
Panel B: Correlation Matrix						
1. Compression	1.00					
2. Coil	.54	1.00				
3. Vacuum	−.35	−.11	1.00			
4. Timing	−.36	−.04	.48	1.00		
5. Alt. & Battery	.17	.18	.11	−.07	1.00	
6. Wheel & Brake	.12	.56	.12	−.15	.25	1.00

Note: The covariance and correlation matrices are based on standardized efficiency scores for each of the duty areas.

TABLE 7 (Continued)

Administrator 4

	Duty Area					
	1	2	3	4	5	6
			Panel A: Covariance Matrix			
1. Compression	111.60					
2. Coil	30.28	172.62				
3. Vacuum	59.90	9.90	163.03			
4. Timing	40.38	10.57	44.93	76.82		
5. Alt. & Battery	−12.03	8.54	31.91	−3.60	99.34	
6. Wheel & Brake	13.26	45.48	15.14	−4.29	8.79	49.91
Mean	48.33	55.05	53.30	51.92	52.81	54.47
Std	10.56	13.14	12.77	8.76	9.97	7.06
N	27	27	27	27	27	27
			Panel B: Correlation Matrix			
1. Compression	1.00					
2. Coil	.22	1.00				
3. Vacuum	.44	.06	1.00			
4. Timing	.44	.09	.40	1.00		
5: Alt. & Battery	−.11	.07	.25	−.04	1.00	
6. Wheel & Brake	.18	.48	.17	−.07	.12	1.00

Note: The covariance and correlation matrices are based on standardized efficiency scores for each of the duty areas.

The matrices for the total sample show that the duty areas are only slightly intercorrelated. The median correlation coefficient is only 0.11, but at least all values are positive. Such is not the case when individual administrators are examined. Administrator 1B has a total of 16 negative correlations, while administrators 2 and 3 each have 12 negative relationships. The other two administrators have a few negative, although insignificant, intercorrelations. These findings imply that the administrators were not applying consistent scoring strategies across the duty areas, as shown by the profiles in figure 8. Note the interaction among the administrators. Their profiles consistently cross, and no parallelism exists. Thus it appears that the scoring rules were inadequately defined and little, if any agreement, concerning the measurement of the job performance construct could be reached by the administrators.

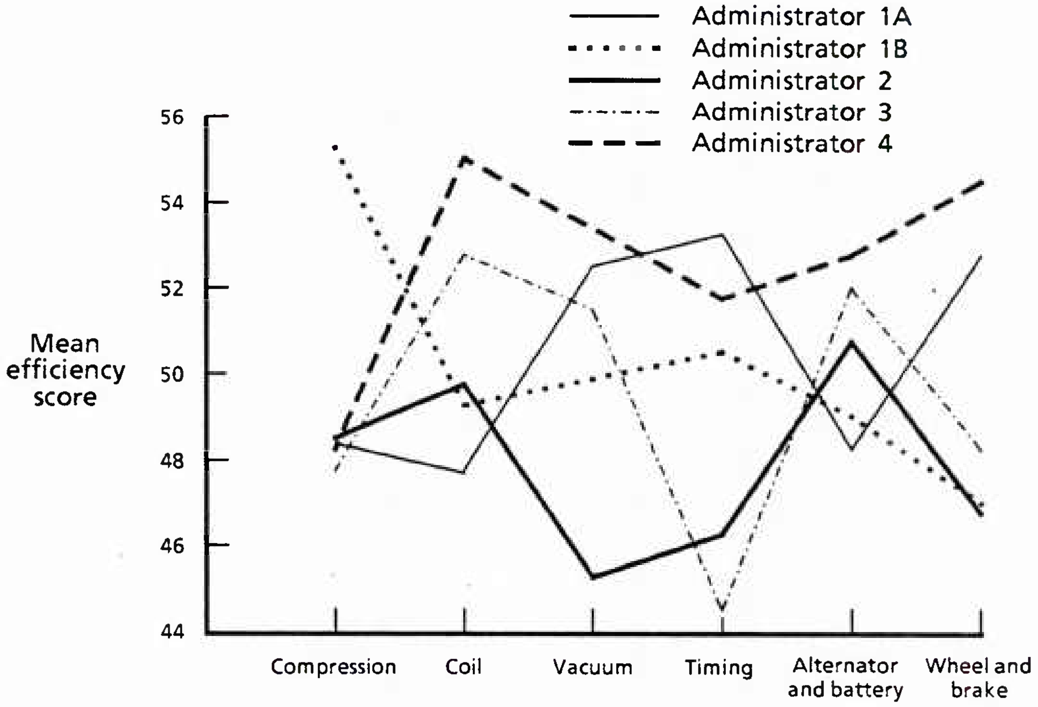

FIG. 8: EQUIVALENCE-OF-ADMINISTRATOR PROFILES FOR AUTOMOTIVE MECHANIC SPECIALTY

The statistical test presented in table 8 confirms that the administrators used different scoring standards. The weighted sum of the determinants of the variance-covariance matrix for each administrator is compared to the determinant for the pooled or total sample in which the administrator variable is disregarded. Note that the determinants of administrators 1B and 2 are the smallest (because of the negative covariances) and therefore these administrators are the most deviant of the group. The test statistic for comparing the determinants is distributed as a chi-square and is highly significant. Therefore a single covariance matrix does not provide the best description of the interrelationships among the data, implying that at least some of the administrators scored the hands-on test differently from others.

TABLE 8

SIGNIFICANCE TESTS FOR EQUIVALENCY AND CONSISTENCY COMPONENTS FOR AUTOMOTIVE MECHANIC SPECIALTY

Administrator	Determinant	Log (Determinant)	N
Pooled	$7.70 * 10^{11}$	27.37	176
1A	$8.87 * 10^{11}$	27.51	55
1B	$5.21 * 10^{10}$	24.68	40
2	$3.06 * 10^{10}$	24.15	41
3	$8.91 * 10^{10}$	25.21	13
4	$4.12 * 10^{11}$	26.75	27
Average	$8.51 * 10^{11}$		

	M	C	df	X^2	Probability
Equivalency test	277.12	.093	84	251.30	< .01
Consistency test	17.06	.014	19	16.83	< .60

Note: The measurement scale is standardized efficiency scores. The pooled administrator represents the total sample, summing over all administrators. The "average" value ($8.5 * 10^{11}$) is based on a matrix with the mean variance and covariance of the pooled sample as its elements. Calculation of the statistics for these tests is described in appendix A.

Given this inconsistency among administrators, steps were taken to edit and/or standardize the Automotive Mechanic data set in order to achieve a uniform set of administrators. The results shown in table 7 already involved some editing in the form of deleting examinees with irregular discharges and three outliers after standardizing the duty areas. The test for equality of administrators was still highly significant as noted in table 8. Applying a log transformation to the efficiency scores and restandardizing and deleting outliers resulted in the deletion of two cases but still did not result in equivalent administrators.

In an effort to control for the administrator-by-duty area interaction (as shown in figure 8), both administrators and duty areas were standardized, and eight more outliers were deleted. But, again, a uniform set of administrators was not identified. The resulting covariance matrices were examined to identify any inconsistent task measurement. No particular tasks were found that unequally contributed to the differences in scoring among the administrators. Likewise there was no administrator whose deletion would result in a more equivalent group. Thus, despite these extreme statistical efforts to yield a set of equivalent administrators, no such group could be identified.

The implication of this finding is that if the process of calibrating administrators is grossly unsuccessful, there is no way to produce a data set of equivalent administrators. This point may appear obvious, but it will almost certainly arise again in the full-scale administration of the Marine Corps JPM Project. Statistical manipulation cannot produce consistent administrators if little or no effort was taken to make them equivalent in the first place. Statistical corrections are best used only for fine tuning and not for gross realignments.

Ground Radio Repair

The covariance and correlation matrices for 135 ground radio repairers are shown in table 9. The matrices are based on proportion-correct scores for each of the ten circuit boards. The boards were not standardized because the scores are on similar scales of measurement. For the total sample, the boards are moderately and positively interrelated. When examining the individual administrators, the most striking results were found for administrator 2. The covariance matrix for this administrator is singular, implying that its determinant is zero. This is due to the scores assigned on board 5 — perfect for

TABLE 9

COVARIANCE AND CORRELATION MATRICES FOR THE CIRCUIT BOARDS OF THE GROUND RADIO REPAIR TEST

Total Sample

	Board									
	1	2	3	4	5	6	7	8	9	10
					Panel A: Covariance Matrix					
Board 1	.04									
Board 2	.02	.06								
Board 3	.02	.03	.09							
Board 4	.01	.03	.03	.10						
Board 5	.01	.00	.03	.03	.06					
Board 6	.00	.01	.02	.02	.02	.07				
Board 7	.02	.01	.02	.03	.02	.01	.09			
Board 8	.01	.02	.04	.04	.04	.02	.05	.14		
Board 9	.01	.02	.03	.04	.03	.02	.02	.04	.09	
Board 10	.02	.03	.04	.05	.04	.02	.04	.07	.06	.14
Mean	.90	.88	.75	.81	.84	.76	.80	.69	.84	.71
Std	.20	.24	.30	.31	.25	.27	.31	.37	.31	.38
N	135	135	135	135	135	135	135	135	135	135
					Panel B: Correlation Matrix					
Board 1	1.00									
Board 2	.40	1.00								
Board 3	.28	.35	1.00							
Board 4	.23	.34	.33	1.00						
Board 5	.13	.08	.39	.35	1.00					
Board 6	.09	.20	.23	.27	.23	1.00				
Board 7	.30	.19	.19	.32	.26	.18	1.00			
Board 8	.13	.21	.37	.37	.39	.23	.46	1.00		
Board 9	.20	.29	.31	.38	.40	.19	.24	.34	1.00	
Board 10	.23	.32	.38	.41	.42	.15	.38	.48	.56	1.00

Note: The covariance and correlation matrices are based on the proportion-correct score for each of the ten circuit boards.

TABLE 9 (Continued)

Administrator Unknown

						Board					
		1	2	3	4	5	6	7	8	9	10
						Panel A: Covariance Matrix					
Board	1	.06									
Board	2	.03	.06								
Board	3	.03	.03	.13							
Board	4	.01	.02	.05	.11						
Board	5	–.00	–.01	.04	.04	.09					
Board	6	.00	.00	.03	.02	.03	.08				
Board	7	.03	.03	.04	.03	.02	.00	.12			
Board	8	.00	.01	.05	.03	.04	.02	.04	.16		
Board	9	.03	.03	.06	.05	.02	.02	.03	.03	.12	
Board	10	.04	.03	.06	.05	.03	.00	.03	.05	.08	.18
Mean		.84	.85	.66	.74	.72	.75	.71	.57	.78	.56
Std		.25	.25	.35	.33	.30	.29	.34	.40	.35	.42
N		46	46	46	46	46	46	46	46	46	46
						Panel B: Correlation Matrix					
Board	1	1.00									
Board	2	.41	1.00								
Board	3	.37	.32	1.00							
Board	4	.16	.19	.43	1.00						
Board	5	–.03	–.15	.41	.40	1.00					
Board	6	.04	.04	.26	.21	.29	1.00				
Board	7	.40	.32	.31	.25	.22	.05	1.00			
Board	8	.03	.06	.35	.21	.34	.18	.30	1.00		
Board	9	.34	.37	.47	.39	.22	.23	.23	.24	1.00	
Board	10	.36	.29	.38	.32	.25	.02	.23	.27	.52	1.00

Note: The covariance and correlation matrices are based on the proportion-correct score for each of the ten circuit boards.

TABLE 9 (Continued)

Administrator 1

	Board									
	1	2	3	4	5	6	7	8	9	10
	Panel A: Covariance Matrix									
Board 1	.05									
Board 2	.03	.10								
Board 3	.01	.04	.09							
Board 4	.01	.07	.01	.12						
Board 5	.00	.04	.02	.03	.05					
Board 6	−.01	.02	.02	.01	.02	.06				
Board 7	−.00	−.00	−.00	.01	.00	.00	.00			
Board 8	−.01	.03	.02	.08	.03	.03	.01	.09		
Board 9	−.01	.04	.03	.09	.04	.03	.01	.09	.10	
Board 10	.00	.06	.04	.04	.04	.04	−.00	.05	.06	.08
Mean	.85	.86	.77	.84	.88	.78	.98	.87	.88	.86
Std	.23	.31	.29	.35	.23	.25	.05	.30	.32	.28
N	16	16	16	16	16	16	16	16	16	16
	Panel B: Correlation Matrix									
Board 1	1.00									
Board 2	.43	1.00								
Board 3	.21	.38	1.00							
Board 4	.19	.68	.09	1.00						
Board 5	.08	.49	.34	.34	1.00					
Board 6	−.20	.26	.28	.12	.33	1.00				
Board 7	−.14	−.18	−.31	.30	.03	.32	1.00			
Board 8	−.19	.30	.18	.72	.44	.39	.38	1.00		
Board 9	−.15	.40	.30	.78	.53	.34	.38	.93	1.00	
Board 10	.00	.65	.53	.42	.67	.60	−.10	.63	.64	1.00

Note: The covariance and correlation matrices are based on the proportion-correct score for each of the ten circuit boards.

TABLE 9 (Continued)

Administrator 2

		Board									
		1	2	3	4	5	6	7	8	9	10
						Panel A: Covariance Matrix					
Board	1	.04									
Board	2	.05	.09								
Board	3	.02	.05	.07							
Board	4	−.00	−.01	−.00	.02						
Board	5	.00	.00	.00	.00	.00					
Board	6	.03	.05	.04	.00	.00	.05				
Board	7	−.00	−.00	−.01	−.00	.00	−.00	.01			
Board	8	.05	.06	.03	.00	.00	.04	−.00	.06		
Board	9	−.00	−.00	.00	.00	.00	−.00	−.00	−.00	.01	
Board	10	−.00	−.00	−.01	.01	.00	−.00	−.00	−.00	−.00	.01
Mean		.95	.89	.78	.93	1.00	.87	.98	.91	.98	.98
Std		.20	.29	.26	.14	.00	.22	.08	.24	.08	.08
N		13	13	13	13	13	13	13	13	13	13
						Panel B: Correlation Matrix					
Board	1	1.00									
Board	2	.91	1.00								
Board	3	.40	.61	1.00							
Board	4	−.14	−.19	−.01	1.00						
Board	5	.00	.00	.00	.00	.00					
Board	6	.60	.81	.62	.03	.00	1.00				
Board	7	−.08	−.11	−.25	−.14	.00	−.18	1.00			
Board	8	.95	.85	.41	.12	.00	.65	−.11	1.00		
Board	9	−.08	−.11	.24	.17	.00	−.18	−.08	−.11	1.00	
Board	10	−.08	−.11	−.25	.48	.00	−.18	−.08	−.11	−.08	1.00

Note: The covariance and correlation matrices are based on the proportion-correct score for each of the ten circuit boards.

TABLE 9 (Continued)

Administrator 3

	Board									
	1	2	3	4	5	6	7	8	9	10
	Panel A: Covariance Matrix									
Board 1	.00									
Board 2	–.00	.00								
Board 3	.00	.01	.04							
Board 4	–.00	.01	.01	.03						
Board 5	–.00	–.00	.01	–.00	.03					
Board 6	.00	.01	.03	.02	.01	.09				
Board 7	–.00	.01	.02	.01	.02	.06	.11			
Board 8	.00	.01	.02	.02	.03	.05	.06	.10		
Board 9	–.00	–.00	–.01	–.00	.03	.00	.00	.03	.09	
Board 10	–.00	–.00	.00	.00	.03	.02	.03	.04	.06	.09
Mean	.98	.98	.80	.91	.92	.74	.76	.74	.90	.79
Std	.05	.07	.21	.18	.17	.30	.33	.32	.29	.31
N	19	19	19	19	19	19	19	19	19	19
	Panel B: Correlation Matrix									
Board 1	1.00									
Board 2	–.08	1.00								
Board 3	.02	.43	1.00							
Board 4	–.04	.65	.33	1.00						
Board 5	–.17	–.12	.17	–.14	1.00					
Board 6	.04	.26	.54	.39	.18	1.00				
Board 7	–.26	.24	.32	.23	.42	.58	1.00			
Board 8	.02	.44	.24	.43	.48	.53	.58	1.00		
Board 9	–.12	–.08	–.10	–.05	.54	.01	.02	.33	1.00	
Board 10	–.07	–.16	.01	.05	.62	.23	.33	.40	.70	1.00

Note: The covariance and correlation matrices are based on the proportion-correct score for each of the ten circuit boards.

TABLE 9 (Continued)

Administrator 4

		Board									
		1	2	3	4	5	6	7	8	9	10
Panel A: Covariance Matrix											
Board	1	.02									
Board	2	.01	.03								
Board	3	−.00	−.00	.06							
Board	4	.03	.01	.01	.06						
Board	5	.02	−.00	.01	.04	.05					
Board	6	.01	.00	.00	.02	.02	.05				
Board	7	.03	.02	.00	.05	.03	.00	.09			
Board	8	.01	.00	.04	.04	.03	−.01	.07	.12		
Board	9	.01	.01	−.01	.02	.03	.02	.02	.01	.06	
Board	10	.00	.01	.03	.02	.02	.02	.05	.07	.02	.09
Mean		.94	.89	.84	.92	.87	.70	.80	.71	.86	.82
Std		.14	.19	.24	.24	.22	.22	.30	.35	.24	.29
N		21	21	21	21	21	21	21	21	21	21
Panel B: Correlation Matrix											
Board	1	1.00									
Board	2	.28	1.00								
Board	3	−.01	−.05	1.00							
Board	4	.89	.18	.11	1.00						
Board	5	.52	−.06	.19	.72	1.00					
Board	6	.29	.08	.04	.44	.45	1.00				
Board	7	.61	.39	.02	.64	.44	.07	1.00			
Board	8	.29	.00	.45	.45	.39	−.09	.67	1.00		
Board	9	.31	.28	−.14	.43	.48	.34	.25	.07	1.00	
Board	10	.02	.26	.42	.26	.38	.23	.51	.64	.31	1.00

Note: The covariance and correlation matrices are based on the proportion-correct score for each of the ten circuit boards.

TABLE 9 (Continued)

Administrator 5

	Board									
	1	2	3	4	5	6	7	8	9	10
					Panel A: Covariance Matrix					
Board 1	.02									
Board 2	.00	.08								
Board 3	−.00	.04	.10							
Board 4	.01	.06	.04	.16						
Board 5	.00	.02	.03	.00	.05					
Board 6	.01	.03	−.00	.06	−.01	.08				
Board 7	.02	.01	−.00	.05	−.01	.02	.11			
Board 8	.01	.04	.05	.04	.01	.01	.05	.14		
Board 9	−.00	.03	.03	.02	.04	.01	.03	.03	.12	
Board 10	.01	.05	.06	.08	.03	.02	.06	.10	.08	.18
Mean	.95	.84	.76	.69	.89	.79	.76	.59	.81	.57
Std	.14	.29	.31	.40	.22	.28	.33	.38	.35	.42
N	20	20	20	20	20	20	20	20	20	20
					Panel B: Correlation Matrix					
Board 1	1.00									
Board 2	.01	1.00								
Board 3	−.06	.47	1.00							
Board 4	.15	.49	.35	1.00						
Board 5	.08	.35	.51	.01	1.00					
Board 6	.26	.36	−.02	.53	−.14	1.00				
Board 7	.43	.13	−.05	.38	−.16	.17	1.00			
Board 8	.16	.35	.42	.29	.14	.07	.42	1.00		
Board 9	−.02	.25	.29	.15	.56	.06	.30	.26	1.00	
Board 10	.13	.44	.44	.48	.38	.20	.46	.60	.56	1.00

Note: The covariance and correlation matrices are based on the proportion-correct score for each of the ten circuit boards.

all 13 examinees. Thus the variance for this board is zero, and accordingly all covariances are indeterminant. This administrator also assigned near-perfect scores with low variances for boards 7, 9, and 10. With this small variance across the ten boards, the correlations among the boards may be spurious. In fact, for administrator 2 there are some 42 negative correlations among the boards. Consequently, this administrator was dropped from the analysis.

The results of the tests for equivalence of administrators are shown in table 10 with administrator 2 deleted. The determinants readily show that administrators 1, 3, and 4 are different from the pooled matrix. The administrator-by-board profile in figure 9 illustrates the deviancy of these administrators, with administrator 1 being extreme for boards 7, 8, and 10, and administrator 3 being the outlier for boards 2 and 1. The deviancy of administrator 4 is not readily identified from the profile.

Note from the profile that the unknown administrators consistently assigned the lowest score. Considering this group to be a baseline since the administrators had no stake in the ratings they assigned, no other administrator even approached the line pattern for the unknown group. In particular, for board 5 the unknown group was in complete opposition to the other four administrators. Thus, there seems to be little consistency among the known administrators and no standard baseline against which to compare their scoring strategies.

The data editing and standardization process discussed earlier to achieve a group of equivalent administrators for the Automotive Mechanic sample was also applied to the Ground Radio Repair sample (except for the log transformation). The matrix resulting from editing the data still showed that the administrators used significantly different scoring standards.

A final step in editing the data was to examine the covariances of each board for inconsistencies with the other boards. Board 1 was identified as having a large number of negative relationships. Thus board 1 was deleted under the assumption that it was mainly a practice item and did not contribute to the examinee's overall score. The covariance matrices for the administrators were then found to be statistically equivalent. The matrices for the total sample and for each administrator after restandardization are presented in table 11. Only a few negative relationships remain.

TABLE 10

SIGNIFICANCE TESTS FOR THE EQUALITY AND CONSISTENCY COMPONENTS FOR THE GROUND RADIO REPAIR SPECIALTY

Administrator	Determinant	Log (Determinant)	N
Pooled	$2.05 * 10^{-12}$	–26.91	122
Unknown	$1.36 * 10^{-11}$	–25.02	46
1	$6.88 * 10^{-18}$	–39.52	16
3	$1.96 * 10^{-17}$	–38.47	19
4	$2.13 * 10^{-16}$	–36.09	21
5	$3.14 * 10^{-13}$	–28.79	20
Average	$5.72 * 10^{-12}$		

	M	C	df	χ^2	Probability
Equality test	564.64	.196	220	453.81	< .01
Consistency test	119.87	.031	53	116.19	< .01

Note: Administrator 2 was deleted because the covariance matrix was singular. The measurement scale is percentage-correct score. The pooled administrator represents the total sample, summing over all administrators. The "average" value ($5.72 * 10^{-12}$) is based on a matrix with the mean variance and covariance of the pooled sample as its elements. Calculation of the statistics for these tests is described in appendix A.

The results of the statistical test in table 12 indicate that an equivalent set of administrators could be formed, but only after deleting both a test administrator and a circuit board—a significant loss of data. Comparing the matrices for the *total sample* before and after this editing process shows that the magnitude of the correlations differentially changed depending upon the relationships among the boards for the second administrator. Of course, on the administrator level, the correlations of the boards were not affected by deleting the second administrator.

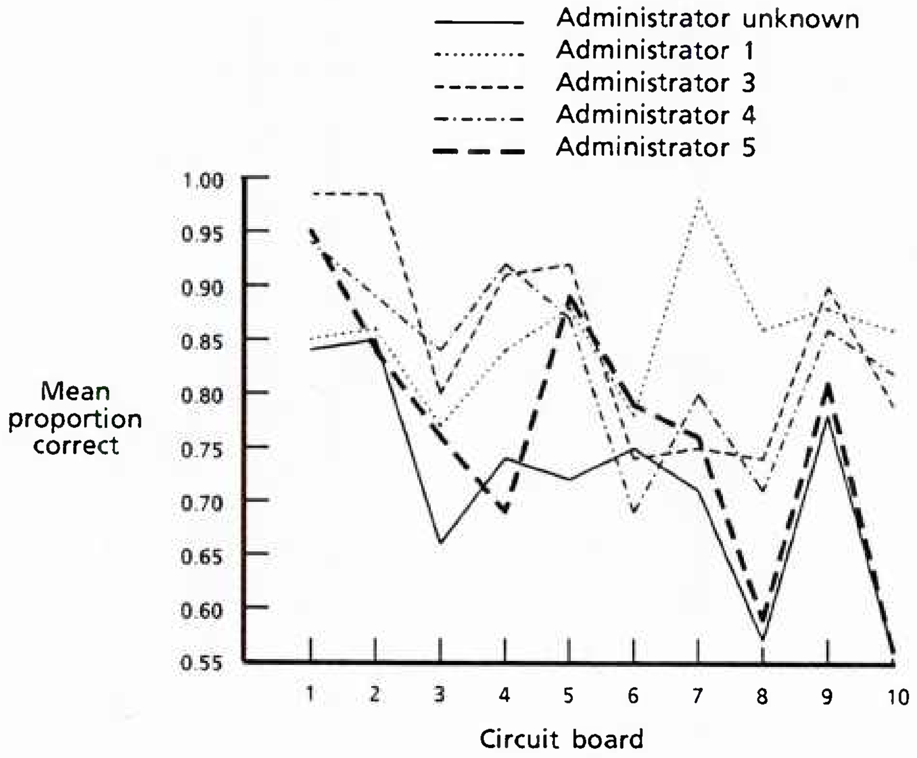

FIG. 9: EQUIVALENCE-OF-ADMINISTRATOR PROFILES FOR GROUND RADIO REPAIR SPECIALTY

Note in table 12 that administrator 1 was the most unlike the other administrators, as shown by the magnitude of his determinant in comparison to the pooled determinant. But this difference was not statistically significant, especially given that he tested only 16 examinees.

The problem of deviant administrators presumably could be reduced with proper training and monitoring of their scoring standards. But the issue of changing the content of the test is more serious. The test has already been judged to represent job requirements, and changing it just to attain desirable statistical properties may degrade content validity.

A lesson to be learned is that practice items should be included on the test so that all examinees begin at the same level of understanding of what they are to do. With practice in taking hands-on tests, the scores should be better measures of individuals' performance and not of their test taking ability.

TABLE 11

COVARIANCE AND CORRELATION MATRICES FOR THE CIRCUIT BOARDS OF THE GROUND RADIO REPAIR TEST HAVING STANDARDIZED ADMINISTRATORS AND BOARDS

Total Sample

	Board								
	2	3	4	5	6	7	8	9	10
				Panel A: Covariance Matrix					
Board 2	96.69								
Board 3	29.46	96.69							
Board 4	35.87	29.25	96.69						
Board 5	3.12	33.28	29.19	96.69					
Board 6	15.46	21.45	30.64	22.77	96.69				
Board 7	21.69	12.17	33.08	19.56	18.23	96.69			
Board 8	18.11	32.96	35.20	33.96	18.95	42.16	96.69		
Board 9	25.79	21.99	33.21	39.52	19.30	22.70	30.74	96.69	
Board 10	27.69	34.62	29.88	39.15	18.66	27.85	43.78	51.49	96.69
Mean	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
Std	9.83	9.83	9.83	9.83	9.83	9.83	9.83	9.83	9.83
N	122	122	122	122	122	122	122	122	122
				Panel B: Correlation Matrix					
Board 2	1.00								
Board 3	.30	1.00							
Board 4	.37	.30	1.00						
Board 5	.03	.34	.30	1.00					
Board 6	.16	.22	.32	.24	1.00				
Board 7	.22	.13	.34	.20	.19	1.00			
Board 8	.19	.34	.36	.35	.20	.44	1.00		
Board 9	.27	.23	.34	.41	.20	.23	.32	1.00	
Board 10	.29	.36	.31	.40	.19	.29	.45	.53	1.00

Note: Administrator 2 and Board 1 have been deleted to achieve a data set of equivalent administrators.

TABLE 11 (Continued)

Administrator Unknown

		Board								
		2	3	4	5	6	7	8	9	10
					Panel A: Covariance Matrix					
Board	2	100.00								
Board	3	31.96	100.00							
Board	4	19.01	42.57	100.00						
Board	5	−15.32	41.16	40.24	100.00					
Board	6	3.92	25.74	20.81	28.89	100.00				
Board	7	31.79	31.12	25.48	21.69	4.72	100.00			
Board	8	6.27	35.39	21.37	34.18	18.05	29.90	100.00		
Board	9	36.53	46.95	39.35	21.74	22.92	23.46	24.03	100.00	
Board	10	29.18	37.93	32.16	25.41	1.89	22.69	26.97	51.50	100.00
Mean		50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
Std		10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
N		46	46	46	46	46	46	46	46	46
					Panel B: Correlation Matrix					
Board	2	1.00								
Board	3	.32	1.00							
Board	4	.19	.43	1.00						
Board	5	−.15	.41	.40	1.00					
Board	6	.04	.26	.21	.29	1.00				
Board	7	.32	.31	.25	.22	.05	1.00			
Board	8	.06	.35	.21	.34	.18	.30	1.00		
Board	9	.37	.47	.39	.22	.23	.23	.24	1.00	
Board	10	.29	.38	.32	.25	.02	.23	.27	.52	1.00

Note: Administrator 2 and Board 1 have been deleted to achieve a data set of equivalent administrators.

TABLE 11 (Continued)

Administrator 1

	Board								
	2	3	4	5	6	7	8	9	10
					Panel A: Covariance Matrix				
Board 2	100.00								
Board 3	38.38	100.00							
Board 4	67.94	9.02	100.00						
Board 5	48.60	34.27	34.29	100.00					
Board 6	26.08	28.41	12.40	33.05	100.00				
Board 7	−17.86	−30.80	30.26	3.00	32.07	100.00			
Board 8	30.25	17.80	71.93	44.35	38.88	38.11	100.00		
Board 9	39.61	30.30	78.41	52.62	33.67	38.33	93.23	100.00	
Board 10	64.99	52.63	42.05	66.64	60.06	−9.93	62.59	63.92	100.00
Mean	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
Std	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
N	16	16	16	16	16	16	16	16	16
					Panel B: Correlation Matrix				
Board 2	1.00								
Board 3	.38	1.00							
Board 4	.68	.09	1.00						
Board 5	.49	.34	.34	1.00					
Board 6	.26	.28	.12	.33	1.00				
Board 7	−.18	−.31	.30	.03	.32	1.00			
Board 8	.30	.18	.72	.44	.39	.38	1.00		
Board 9	.40	.30	.78	.53	.34	.38	.93	1.00	
Board 10	.65	.53	.42	.67	.60	−.10	.63	.64	1.00

Note: Administrator 2 and Board 1 have been deleted to achieve a data set of equivalent administrators.

TABLE 11 (Continued)

Administrator 3

Panel A: Covariance Matrix

		Board								
		2	3	4	5	6	7	8	9	10
Board	2	100.00								
Board	3	42.63	100.00							
Board	4	64.92	33.04	100.00						
Board	5	−11.52	16.74	−14.3	100.00					
Board	6	25.53	53.68	39.06	18.11	100.00				
Board	7	24.27	32.05	22.56	42.45	58.17	100.00			
Board	8	44.41	23.86	42.70	47.82	53.01	58.38	100.00		
Board	9	−8.06	−10.18	−4.75	54.36	.95	2.46	33.43	100.00	
Board	10	−16.47	.75	5.44	62.06	23.26	33.35	40.46	70.30	100.00
Mean		50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
Std		10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
N		19	19	19	19	19	19	19	19	19

Panel B: Correlation Matrix

		2	3	4	5	6	7	8	9	10
Board	2	1.00								
Board	3	.43	1.00							
Board	4	.65	.33	1.00						
Board	5	−.12	.17	−.14	1.00					
Board	6	.26	.54	.39	.18	1.00				
Board	7	.24	.32	.23	.42	.58	1.00			
Board	8	.44	.24	.43	.48	.53	.58	1.00		
Board	9	−.08	−.10	−.05	.54	.01	.02	.33	1.00	
Board	10	−.16	.01	.05	.62	.23	.33	.40	.70	1.00

Note: Administrator 2 and Board 1 have been deleted to achieve a data set of equivalent administrators.

TABLE 11 (Continued)

Administrator 4

	Board								
	2	3	4	5	6	7	8	9	10
				Panel A: Covariance Matrix					
Board 2	100.00								
Board 3	−5.14	100.00							
Board 4	18.20	11.06	100.00						
Board 5	−6.16	19.44	72.34	100.00					
Board 6	7.69	4.04	43.89	45.40	100.00				
Board 7	38.51	2.37	63.98	44.20	6.77	100.00			
Board 8	.00	44.95	45.11	39.28	−9.15	67.12	100.00		
Board 9	27.58	−14.03	43.19	48.32	33.77	25.48	7.26	100.00	
Board 10	25.79	42.16	26.11	37.73	23.28	50.77	63.90	31.20	100.00
Mean	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
Std	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
N	21	21	21	21	21	21	21	21	21
				Panel B: Correlation Matrix					
Board 2	1.00								
Board 3	−.05	1.00							
Board 4	.18	.11	1.00						
Board 5	−.06	.19	.72	1.00					
Board 6	.08	.04	.44	.45	1.00				
Board 7	.39	.02	.64	.44	.07	1.00			
Board 8	.00	.45	.45	.39	−.09	.67	1.00		
Board 9	.28	−.14	.43	.48	.34	.25	.07	1.00	
Board 10	.26	.42	.26	.38	.23	.51	.64	.31	1.00

Note: Administrator 2 and Board 1 have been deleted to achieve a data set of equivalent administrators.

TABLE 11 (Continued)

Administrator 5

		Board								
		2	3	4	5	6	7	8	9	10
		Panel A: Covariance Matrix								
Board	2	100.00								
Board	3	46.63	100.00							
Board	4	49.15	35.39	100.00						
Board	5	35.19	51.09	1.00	100.00					
Board	6	36.33	−1.89	52.82	−14.42	100.00				
Board	7	13.36	−4.76	37.69	−15.91	17.33	100.00			
Board	8	34.50	42.10	28.85	13.64	6.65	41.64	100.00		
Board	9	25.05	29.19	15.41	56.31	5.60	29.58	25.96	100.00	
Board	10	44.26	43.98	48.27	38.02	20.39	46.41	59.91	56.03	100.00
Mean		50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
Std		10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
N		20	20	20	20	20	20	20	20	20
		Panel B: Correlation Matrix								
Board	2	1.00								
Board	3	.47	1.00							
Board	4	.49	.35	1.00						
Board	5	.35	.51	.01	1.00					
Board	6	.36	−.02	.53	−.14	1.00				
Board	7	.13	−.05	.38	−.16	.17	1.00			
Board	8	.35	.42	.29	.14	.07	.42	1.00		
Board	9	.25	.29	.15	.56	.06	.30	.26	1.00	
Board	10	.44	.44	.48	.38	.20	.46	.60	.56	1.00

Note: Administrator 2 and Board 1 have been deleted to achieve a data set of equivalent administrators.

TABLE 12

SIGNIFICANCE TESTS FOR THE EQUIVALENCY AND CONSISTENCY COMPONENTS FOR THE GROUND RADIO REPAIR SPECIALTY (STANDARDIZED ADMINISTRATORS AND CIRCUIT BOARDS)

Administrator	Determinant	Log (Determinant)	N
Pooled	$9.58 * 10^{16}$	39.10	122
Unknown	$1.21 * 10^{17}$	39.34	46
1	$5.02 * 10^{13}$	31.54	16
3	$8.53 * 10^{15}$	36.68	19
4	$6.50 * 10^{15}$	36.41	21
5	$2.11 * 10^{16}$	37.59	20
Average	$1.61 * 10^{17}$		

	M	C	df	X^2	Probability
Equality test	242.69	.179	180	199.67	.15
Consistency test	60.73	.032	43	58.80	.05

Note: Administrator 2 was deleted because the covariance matrix was singular, and Board 1 was dropped because of its inconsistency with other boards. The measurement scale is percentage-correct score standardized across administrators and boards. The pooled administrator represents the total sample, summing over all administrators. The "average" value ($1.61 * 10^{17}$) is based on a matrix with the mean variance and covariance as its elements. Calculation of the statistics for these tests is described in appendix A.

CONSISTENCY OF TASK MEASUREMENT

Even though job requirements in an MOS are diverse, the intent of hands-on testing is to generalize from the test scores to proficiency in the MOS. The correlations of scores among the tasks in the hands-on test, therefore, should be positive.

For the Automotive Mechanic specialty, the correlations among the tasks were presented earlier in table 7. The correlations are all positive, but they range from only 0.02 to 0.27. Recall that the equivalency of the administrators for this specialty could not be established (see table 8). That confounds any effort to determine the consistency of task measurement for the Automotive Mechanic test.

For the Ground Radio Repair specialty, the correlations of scores among the circuit boards are all positive, with the average correlation being almost 0.30. This implies that all the boards are generally consistent measures of job proficiency.

Table 13 presents the correlations among the 12 tasks for the Infantry Rifleman hands-on test. The "firing upon friendly targets" task is reversely scored, with low scores implying higher performance. Thus the negative relationships between this scale and the scores of other tasks are expected. Note that essentially all correlations are positive (or in the expected direction) as required for the consistency of task measurement assumption. Correlations among the tasks scores are generally much higher within a duty area than across all duty areas. The exception is the target engagement duty area.

STANDARDIZATION OF TESTING PROCEDURES

There are no statistical tests or specific data from the tests given during the feasibility study that address the question of standardized test administration procedures and their influence on the observed hands-on test scores. There was no report of randomization being used in selecting, assigning, or rotating examinees or administrators. Without such randomization, the results of the feasibility study are limited in their generalizability, and the influence of nuisance variables was not adequately controlled.

TABLE 13

CORRELATIONS AMONG TASK SCORES ON THE INFANTRY RIFLEMAN HANDS-ON TEST

		Task											
	Task	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	Target	1.00											
2	Friend	.06	1.00										
3	Stomach	.03	.03	1.00									
4	Jaw	.03	−.06	.21	1.00								
5	Ableed	.05	.03	.36	.13	1.00							
6	Map	.10	−.08	.13	.14	.09	1.00						
7	Compass	.03	−.14	.05	.13	.09	.44	1.00					
8	Terrain	.15	−.10	.09	.19	.10	.63	.44	1.00				
9	Symbols	.00	−.13	.04	.03	.06	.33	.22	.29	1.00			
10	Situatns	.09	−.09	.10	.03	.03	.33	.15	.28	.47	1.00		
11	Removemn	.23	−.02	.00	.14	.07	.28	.22	.30	.07	.15	1.00	
12	Armmine	.12	−.01	.14	.12	.23	.24	.22	.27	.00	.12	.31	1.00
	Mean	.55	.16	.51	.50	.59	.50	.55	.42	.59	.52	.44	.58
	Std	.23	.20	.21	.34	.18	.23	.38	.18	.22	.25	.24	.27
	N	259	259	259	259	259	259	259	259	259	259	259	259

NOTE: The correlation matrix and descriptive statistics are based on proportion-correct scores. The firing upon friendly targets scale is reversely scored. The meaning of the task labels and the respective duty area (in caps) for each task are as follows:

TARGET ENGAGEMENT		
	Target:	Target firing
	Friend:	Firing upon friendly targets
FIRST AID		
	Stomach:	Stomach wound
	Jaw:	Jaw wound
	Ableed:	Arterial bleeding
MAP AND COMPASS		
	Map:	Use of map
	Compass:	Use of compass
	Terrain:	Identify terrain features
FIRE TEAM FORMATION		
	Symbols:	Identify symbols
	Situatns:	Reaction in specific situations
ANTITANK MINES		
	Removemn:	Disarm and remove mine
	Armmine:	Arm mine

SECTION 3

IMPLICATIONS

SUMMARY OF RESULTS

The proper selection of test content is critical in establishing the measurement validity of hands-on tests. For the three MOSs investigated here, the job experts did not adequately sample the full range of job requirements within the MOS. This deficiency was partially due to the lack of an explicit table of specifications that would outline the scope and emphasis of MOS job requirements and provide a justifiable basis for proper sampling of test content. An important ingredient for developing such a table of specifications is an authoritative description the job requirements, such as those found in the Individual Training Standards (ITS), MOS Manuals, task lists, etc. Unfortunately, the ITS were not available at the time of test development, and the experts had to rely on less definitive descriptions of job requirements.

In addition to not being representative, the hands-on tests were plagued by inappropriate levels of task difficulty. For two specialties, the hands-on tests provided little useful information with respect to individual differences at the top end of the measurement scale because most examinees received nearly perfect marks. To counter this ceiling effect, efficiency scores were used for Automotive Mechanic specialty.

The analyses to evaluate the equivalence of test administrators was conducted for only two of the specialties, Ground Radio Repair and Automotive Mechanic. In neither case were the administrators found to be equivalent raters of performance. However, it should be noted that they received no training in how to score the tests, nor were they monitored or given feedback on the accuracy of their scoring. For the Automotive Mechanic specialty, even rather severe data editing and standarization procedures could not produce a group of administrators who applied equivalent scoring standards. Such data manipulation was successful for the Ground Radio Repair specialty, but only after one extreme administrator was deleted and the first circuit board was dropped from the test.

Consistency of task measurement, as defined by positive intercorrelations of the tasks, was found for each of the specialties. Such consistency was noted on the total sample level only. On the administrator level for the Automotive Mechanic and Infantry Rifleman specialties, essentially zero and some negative task intercorrelations were found. These findings are difficult to interpret because they are confounded by inconsistent scoring among administrators and do not necessarily imply inconsistent measurement. The greatest degree of consistency in task measurement was noted for the circuit boards of the Ground Radio Repair test, as would be expected given the diversity of tasks for the other specialties.

Little, if any, randomization was employed in designing the feasibility study. This lack of randomization may partially contribute to the inconsistencies and anomalies noted in the measurement validity for each of the specialties. However, no explicit tests exist for quantifying this factor.

IMPLICATIONS FOR THE MARINE CORPS JOB PERFORMANCE MEASUREMENT PROJECT

The results of this study are enlightening with respect to developing and administering future hands-on tests. Its implications are addressed below from the perspective of the four components of measurement validity: content validity, equivalence of administrators, consistency of tasks, and standardization of testing procedures.

Content Validity

The most striking implication resulting from the feasibility study is that selection of test content must be based on a detailed table of specifications. Each of the tests for the three MOSs lacked sufficient content validity because the sampling of tasks was inadequate to measure the full range of job requirements. A table of specifications provides such guidance.

In addition, task difficulty should be incorporated into the table of specifications. For the Automotive Mechanic specialty, the proportion-correct scores for the hands-on test were exceedingly high, allowing for little inference to be drawn from individual differences in test performance. Similar results, although not as high, were found for the Ground Radio Repair test. Thus, the job experts did not select tasks for the test that represented the full range of difficulty for the MOS.

The results of the Automotive Mechanic specialty also point to the importance of scaling issues in hands-on testing. While the proportion-correct scale was converted to an efficiency scale that resulted in a better distribution of test scores, the overall emphasis of the hands-on test was significantly changed. Should automotive mechanics be expected to perform to a high standard without regard to time or to perform to a high standard in a short period of time? Such a question, along with other relevant scaling issues, should be addressed by Marine Corps job experts in the beginning of the test development process, and the answers should not be guided by the shape of score distributions or other factors irrelevant to job content. Again, this points to the considerable importance of explicitly defining in detail the skills and knowledge required for performing the tasks of the MOS, in addition to the content areas that define the job requirements.

In developing a proper hands-on test, the parts or subscales of the test need to be as good as the whole. Test content should reflect the weightings that the tasks have within the entire MOS. Recall that the wheel-and-brake subtest for the automotive mechanic specialty required 60 minutes to complete (or one quarter of the total testing time). During this hour, only 13 steps were scored, a rather inefficient use of testing time. Conversely, the coil subtest, which was only one of several in the engine tune-up duty area, had 18 steps scored and required less than 30 minutes.

The total hands-on score resulting from the sum of these and the other subscales has important policy implications for inferring examinees' proficiency in the Automotive Mechanic specialty. This simple summation of unweighted subtest scores implies that proficiency on the coil section is more important than proficiency in the wheel-and-brake area. Thus, the scaling (or lack thereof) of the subscales of the hands-on test has important implications for generalizing from performance on the test to proficiency in the specialty. Just as administrators cannot be assumed to accurately score performance to the same scale, the subscales cannot be assumed to be mere additive products that automatically reflect job proficiency. Precautions about scaling must be taken in the beginning stages of the project. Scores resulting from the subscales of the hands-on test should be scaled to reflect the overall emphasis of that content area within the MOS.

Scores that have decidedly different scales of measurement can be statistically equated before a total test score is computed. For example, the Infantry Rifleman target score ranged from 0 to 92. Similarly, one of the compass tasks had scores ranging from 0 to only 4. Standard deviations of these two scales are significantly different. If the tasks are considered equally important with respect to performance in the MOS, simple adding will not result in a test score that reflects performance in the MOS. Thus, test developers need to consider such scaling issues and define the appropriate level at which tasks are thought to be equivalent to justify subtest score summation.

Another content validity consideration centers on the primary purpose of the testing. The hands-on test for the Ground Radio Repair specialty was developed with the explicit purpose of being the best single *predictor* of the examinees' on-the-job proficiency. This was the rationale for selecting only a single content area (circuit boards) and a single skill and knowledge area (troubleshooting) because these areas were thought to be most predictive of performance in ground radio repair. However, this purpose is in contrast to the purpose of the Marine Corps JPM Project which is to *measure* job proficiency. Establishing the validity of tests intended for these two different purposes requires two different kinds of analysis. The purpose of prediction requires establishing the criterion-related or predictive validity of the test, or minimizing the errors of estimate. Conversely, the measurement purpose necessitates establishing the content validity of the test and minimizing the errors of measurement, as shown in this paper. Thus, it is necessary that the purpose of testing be explicitly stated as measuring job proficiency, and that the selection of test content and follow-up analyses follow from this purpose.

A final issue focuses on the materials used to define the job content and skills and knowledge necessary to complete the job requirements of a given MOSs. In the Marine Corps, ITS are the most comprehensive source of this type of information, but they are not available for all MOSs. However, even when ITS are available, job experts should draw on as broad a base of information as possible in developing the table of specifications. It is essential to use more than just task titles, for titles can often be ambiguous, as in the case of the Infantry Rifleman specialty. While there was reasonable overlap of the test content with the ITS, the crosswalk between the two was not straightforward. Does a map-and-compass subtest reflect the tactical measures duty area or the land navigation duty area? Detailed descriptions of the job are needed to ensure the proper definition of a table of specifications and to estimate the generalizability of the test.

Equivalency of Administrators

The greatest shortcoming of the data for the Infantry Rifleman specialty was the lack of test administrator identification for each examinee.

Administrator effects on the observed hands-on test scores could not be independently estimated. Because, administrators did not rotate among the testing stations, the variance attributable to testing stations is confounded with administrator and duty area. The profile analyses used to monitor test administrators require rotation of administrators. Thus, it is imperative that administrators be identified for each examinee and that they rotate among the testing stations.

The policy change that occured for the Radio Repair specialty during the feasibility study requiring test administrators to sign answer sheets for each person they tested also had implications for future hands-on testing. As administrators became more personally accountable for the scores they assigned, they also became more lenient. Therefore, administrators should be made aware of the necessity of consistent scoring standards and how shifts in their scoring strategies can influence the overall research findings. The profile analyses described in section 1 should help administrators identify and avoid these personal shifts.

The results for the Automotive Mechanic specialty illustrate the need for proper training and monitoring of test administrators. Despite rather extreme steps to edit and standardize the data set, it was not possible to identify a group of administrators whose scoring standards were equivalent. Thus, efforts to calibrate administrators to the same score scale are essential because statistical manipulation cannot always produce a homogenous set of performance raters. Equivalence of test administrators should be built into the project in the beginning by adequate training and monitoring of administrators, with statistical correction of the data being used only as a last resort.

Finally, administrators must get constructive feedback on their scoring. They cannot be expected to use consistent scoring strategies if they do not know how their scoring compares to that of other administrators or to their past scoring patterns. The profiles discussed earlier address these feedback concerns. Using such profiles for correcting deviant administrators and reinforcing consistent administrators requires explicit managerial actions. Periodic debriefing sessions in which test administrators review their profiles and discuss their testing experiences and problems may facilitate equivalence of administrators.

Consistency of Tasks

One practical implication of the findings for the Ground Radio Repair specialty is the possibility of requiring practice tasks (or items) for different duty areas. In the subtest involving the first circuit board, the scores of ground radio repairers showed extremely inconsistent correlations with scores on other boards. This piece of equipment was new to them, and their scores on it probably reflected a great deal of learning. The degree of learning on the first test item of duty areas for other specialties is unknown. However, if the tasks are presented under conditions that are unfamiliar to the examinee, it is likely that scores on the initial test item will not be consistent with the rest of the items for that duty area. Thus, in an effort to ensure that all examinees are at the same level of understanding concerning what is being tested, it is recommended that practice test items be presented for selected duty areas that use different equipment or conditions that are not familiar to the examinees.

Consistency of the tasks should be evaluated during the tryout of hands-on tests. If a task is found to be inconsistent with the others, as evidenced by a pattern of negative correlation coefficients, it should be reviewed by job experts. Perhaps an adjustment to the scoring rules or slight change in content can make the task consistent with the others. If not, the experts need to decide whether in fact the task is a legitimate job requirement; if so, it may stay in the test.

The results of the analysis investigating the consistency of task measurement have serious implications for the interpretation of statistical findings concerning hands-on tests. An important distinction needs to be drawn between statistical significance and practical significance. Recall that the requirement for task symmetry was a highly restrictive assumption. With respect to the essentially parallel circuit boards, the statistical test for task symmetry found that the subtest scores for the various circuit boards were not sufficiently intercorrelated. Thus, in a strict statistical sense, the board subtests are not parallel measures. If each task is treated as an independent measure (as when computing test reliability), then the symmetry assumption is required. However, if the tasks are summed to result in a single dependent measure, then only consistency among the tasks is required, and the symmetry assumption is not required.

Standardization of Testing Procedures

Standardization of testing procedures is necessary to ensure that test scores reflect individuals' true proficiency and not irrelevant variables of the measurement process. Such standardization includes consistency of testing instructions, testing time, administrator assistance, and application of scoring rules. Therefore, standardization is more of a test management function and not a property of the test itself.

Randomization is an important aspect of test standardization. As noted earlier, little if any randomization was incorporated into the feasibility study. However, it should be introduced in the design at every feasible level. Examinees should be randomly selected from the population to reduce the potential influence of nuisance variables as well as ensure generalization of the findings to the entire Infantry specialty. To facilitate random rotation of the examinees among the testing stations and to coordinate the efficient use of the available testing time, a daily testing plan should be developed. By randomly assigning each examinee to a specific path through the stations, the effects of selection, order, and rotation will be minimized. Administrators should be randomly assigned to their initial testing stations and then they should be randomly rotated through all of the stations. Tasks should be randomly selected from a pool of equivalent items and randomly assigned to a test form. These are only a few aspects of the design that should employ randomization. At any point that a seemingly arbitrary decision is to be made, randomization should enter into that decision.

Another aspect of standardization of testing procedures involves the initial tryout of the hands-on test. From the feasibility study, it is evident that the tryout of hands-on tests needs to be more extensive than pretesting five or six individuals. Results based upon such limited sample are not sufficiently stable to be used in evaluating the test materials. Tryouts should be iterative, with adjustments made to improve the testing procedures. Any changes should be thoroughly evaluated before full implementation. In the feasibility study, the tryout analyses were not sensitive enough to detect the ceiling effects or administrator inconsistencies for the automotive mechanics.

It is also recommended that the *same* test administrators be used for the tryout and full-scale testing. Participation in the tryout is a necessary part of the in-depth training of administrators. It is a time when administrators are required to perform under true testing conditions, and their performance is immediately evaluated.

Another implication for standardized testing procedures concerns the rank of the examinee, which might influence the scoring of some administrators. If feasible, the examinees should not wear rank insignia during the testing. Given that the administrators will be well versed in Marine Corps operations, rank may influence the assigning of test scores and thereby artificially reduce the measurement validity of the hands-on test.

Components of Hands-on Test Scores

The thrust of this paper has been to examine the requisite components of hands-on measurement validity. Particular forethought has been given to the implications for the Marine Corps part of the Joint Service job performance effort beginning in 1986 within the Infantry Occupational Field.

The following general linear model describes the potential sources of error that should be tested for the full-scale administration of the hands-on tests for the Infantry Occupational Field:

$$X = E + DA + T(DA) + A + DA * A + E * T(DA) + e$$

where

X = hands-on test score

E = examinee

DA = duty area or testing station

$T(DA)$ = task within duty area

A = administrator

$DA * A$ = duty-area-by-administrator interaction

$E * T(DA)$ = examinee-by-task interaction

e = error.

This model addresses all of the measurement validity concerns in this paper except for the representativeness of the test content, for which there are no statistical tests. It can be used to test for the extent to which these potential sources of error affect observed hands-on test scores. Ideally, most of these factors and their interactions will be insignificant. If not, the model will assist in estimating the magnitude of statistical adjustment necessary to counter such inconsistency. Of particular interest are the DA * A interactions. These interactions should be insignificant, as they are the specific focus of the quality-control efforts that involved drawing interaction profiles to provide feedback to administrators on their scoring.

CONCLUSIONS

- Establishing the measurement validity of the hands-on tests is a critical prerequisite to the larger goal of using ASVAB scores to predict job performance.

- Detailed tables of specifications based on Individual Training Standards and other descriptions of job requirements are needed. They should incorporate three dimensions: content areas, skills and knowledge, and difficulty-to-learn ratings. These tables of specifications will help ensure that the hands-on tests are representative of MOS job requirements.

- Equivalency of administrators cannot be taken for granted, despite the raters' level of expertise in the subject matter. Intensive training and continuous monitoring of test administrators are crucial to the success of the Marine Corps Job Performance Measurement Project.

- Consistency of task measurement of the hands-on test is necessary if a single score of job proficiency is to be computed. Scaling issues need to be addressed in the initial stages of development so that the total test score is formed from essentially equivalent or properly weighted subscales.

- Standardization of testing procedures and conditions is required to ensure that hands-on performance scores are a function of individuals' proficiency and not the extraneous factors of the measurement process. Randomization also needs to be introduced into the data collection to minimize systematic error and to enhance the generalizability of the research findings.

REFERENCES

[1] CNA, Report 89, *An Evaluation of Using Job Performance Tests To Validate ASVAB Qualification Standards*, by Milton H. Maier and Catherine M. Hiatt, Unclassified, May 1984

[2] Air Force Human Resources Laboratory, Special Report 84-26, *Occupational Learning Difficulty: A Standard for Determining the Order of Aptitude Requirement Minimums*, by Joseph Weeks, Unclassified, 1984

[3] Naval Personnel Research and Development Center, NPRDC TN 82-20, *Marine Corps Job Performance Test for Three Enlisted Specialties*, by David J. Chesler, Chester R. Bilinski, and Marc A. Hamovitch, Unclassified, 1982

APPENDIX A

TESTING THE EQUIVALENCE AND CONSISTENCY OF COVARIANCE MATRICES AND RESULTING IMPLICATIONS FOR THE F TEST

APPENDIX A

TESTING THE EQUIVALENCE AND CONSISTENCY OF COVARIANCE MATRICES AND RESULTING IMPLICATIONS FOR THE F TEST

Assume that p administrators give q measures to n_i subjects. In testing the equality of administrators, one seeks to determine if the covariance matrices of the administrators (S_1, S_2, ..., S_p) are random samples from populations in which the covariance matrices are equal ($\Sigma_1 = \Sigma_2 = ... = \Sigma_p$).

The test statistic for the equality assumption is distributed as a chi-square with df_1 degrees of freedom:

$$X^2 = (1 - C_1)\,M_1$$

where

$$M_1 = N\,ln|S_{pooled}| - \Sigma n_i\,ln|S_i|\,.$$

$$C_1 = \frac{2q^2 + 3q - 1}{6\,(q+1)(p-1)}\left[\Sigma\left(\frac{1}{n_i}\right) - \frac{1}{N}\right]$$

$$df_1 = \frac{q(q+1)(p-1)}{2}\ .$$

Rejection of this test implies that the covariance matrices cannot be pooled, and the relationships within the data are better represented by separate covariance matrices for each administrator or some subset of administrators. However, if the populations have a common covariance matrix Σ, then S_{pooled} is an unbiased estimate of Σ.

In addition to the equality assumption, the valid model for the F test in a repeated measures design requires consistency (or symmetry) of the covariance matrix, i.e., equal variances and covariances. If the covariance matrix has symmetric form, then a matrix (represented as S_0) with the mean variance in the diagonal and mean covariance in all off-diagonal elements is an unbiased estimate of Σ. The test statistic for the consistency assumption is distributed as a chi-square with df_2 degrees of freedom:

$$X^2 = (1 - C_2)\, M_2$$

where

$$M_2 = -(N-p)\, ln\, \frac{|S_{pooled}|}{|S_o|}$$

$$C_2 = \frac{q(q+1)^2(2q-3)}{6\,(N-p)\,(q-1)\,(q^2+q-4)}$$

$$df_2 = \frac{q^2+q-4}{2}\ .$$

If this test is significant, then the q measures are not considered to be parallel.

If the chi-square statistics for both the equality and consistency tests are not significant, then both assumptions are satisfied and the usual F tests are valid. However, the implication of nonequivalent and/or nonsymmetric covariances in a repeated measures design is that the F test is no longer an exact test but rather is positively biased. In other words, significant main effects and interactions are found more often than is truly the case.

A conservative F test exists such that the degrees of freedom are adjusted to counter this positive bias. Computational procedures for the conservative F test are identical to those of the conventional F test but with degrees of freedom equal to 1 and $n - 1$ for the repeated measure and error, respectively. This is in contrast to $q - 1$ and $(n - 1)(q - 1)$ for the conventional F test, where q equals the number of repeated measures and n equals the number of subjects.

Figure A-1 shows the relationship between the critical values for the conventional and conservative F tests and three regions that represent areas of significance and nonsignificance for both tests. If the conservative F test for the repeated measure is significant (region A), the exact test will also be significant. If, however, the conventional test is not significant (region B), one can conclude that there is no effect and stop the analysis because the conservative test would likewise not be significant. The problem arises when the conservative test is insignificant but the conventional test is significant (region C). The width of this unknown region between the critical values for the two tests is determined by the degree of heterogeneity of covariances among the tasks; the greater the heterogeneity, the greater the region. Results falling into the ambiguous region imply that the degree of heterogeneity is significant and therefore the tasks are measuring different concepts. At this point, the model for univariate analysis of variance is no longer appropriate, and one must employ a multivariate significance test. Further discussion of these statistical tests and the implications for F test are presented in [A-1].

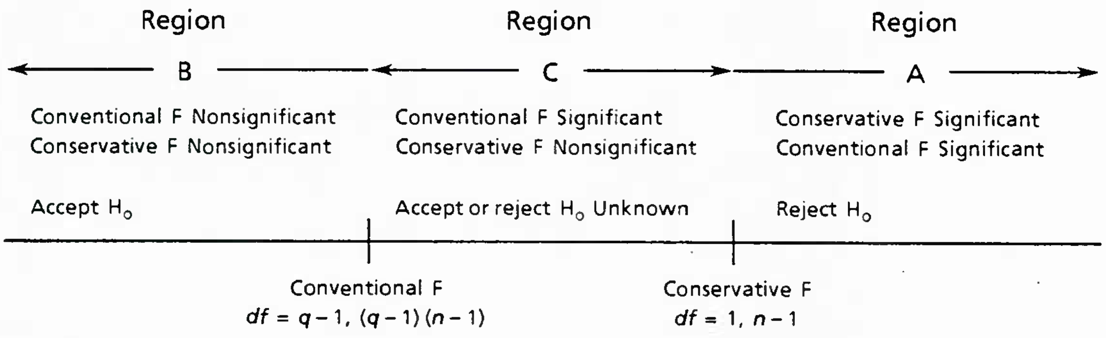
Region
B
Conventional F Nonsignificant
Conservative F Nonsignificant
Accept H_0
Region
C
Conventional F Significant
Conservative F Nonsignificant
Accept or reject H_0 Unknown
Region
A
Conservative F Significant
Conventional F Significant
Reject H_0
Conventional F
$df = q - 1, (q - 1)(n - 1)$
Conservative F
$df = 1, n - 1$

FIG. A-1: RELATIONSHIP BETWEEN CRITICAL VALUES FOR THE CONVENTIONAL AND CONSERVATIVE F STATISTICS

REFERENCE

[A-1] Winer, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill, 1962